Scientific Progress in Artificial Intelligence: History, Status, and Futures

Eric Horvitz and Tom M. Mitchell

Introduction and Background

Artificial Intelligence (AI) refers to a field of endeavor as well as a constellation of technologies. The Association for the Advancement of AI (AAAI) defines the field as pursuing "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines." AI encompasses the development of methods for learning from data, representing knowledge, and performing reasoning aimed at building computer systems capable of performing tasks that typically have required human intelligence. Core capabilities covered in AI research include methods for learning, reasoning, problem-solving, planning, language understanding, and visual perception. Over the last twenty years, AI research transitioned from a niche scientific endeavor to an impactful set of technologies. We provide in this overview chapter a brief history of the evolution of AI as a discipline over nearly seven decades. Then, we review recent advances and directions. This arc through history, present, and the expected near future was commissioned to provide a February 2024 snapshot of the state of AI in support of a series of meetings on AI and the sciences that was organized by the National Academy of Sciences and the Annenberg Trust.

Birth and Evolution of Scientific Field

The prospect of automating aspects of human thinking via mechanical systems has been considered for hundreds of years. Modern metaphors and framing of thinking as a computational process have roots in the early twentieth century. Key contributions to the perspective of thinking as computing include the theoretical work of Alan Turing on computability,[1] efforts by John von Neuman, Turing, and others to construct general-purpose computing systems,[2] and work on computational abstractions of neuronal systems by McCollough and Pitts.[3] The 1940s saw the rise of discussions and publications viewing the computer as a metaphor for the brain, including control-theoretic notions referred to as *cybernetics*.[4]

The modern discipline of AI, per the establishment of a long-standing set of aspirations, harkens back to a research project proposal for a summer workshop held at Dartmouth College in 1956.[5] The proposal, coauthored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, outlined a new field of studying how machines could be programmed to perform "every aspect of learning or any other feature of intelligence." Containing the first use of the phrase *artificial intelligence*, the proposal described goals of finding "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." The summer study is considered as the formal launch of AI as a distinct field of scientific inquiry, setting the foundation for decades of research in computer science.

The maturation of the AI research program saw the evolution of a set of AI subdisciplines with overlapping but distinct research communities, including natural language understanding, problem-solving, planning, vision, robotics, and machine learning. Research areas and communities also formed around distinct foundational approaches to building AI, such as logical reasoning and representations, reasoning under uncertainty with statistical methods, and the use of neural network models versus high-level symbols—a domain of research that had been referred to for decades as *connectionist* approaches. Further, advances and questions in AI have stimulated efforts in other disciplines, such as cognitive psychology, where *cognitive science* refers to a subdiscipline of both AI and cognitive

psychology that centers on taking inspiration from studies, data, and questions about human cognition to build systems that can perform automated learning and reasoning, and on using computational approaches to modeling and probing human psychological processes.[6]

Representations and Reasoning Mechanisms

Scientific studies of AI are best understood in terms of the technical evolution of different approaches to representing and reasoning with data and knowledge. In the early days of the field, representations and reasoning methods included the use of neural networks, early-on referred to as *perceptrons* in work on learning to recognize visual patterns,[7] and symbolic logic applied in both specific instances and in attempts to build general architectures for problem-solving.[8] Symbolic representations dominated the first several decades of AI research with efforts in statistical methods, including neural networks, continuing but largely taking a backstage position. Work in logic-based systems included *rule-based expert systems* that came to focus of attention in the 1970s and 1980s. These systems were aimed at capturing specialist knowledge in sets of compact logical rules (e.g., if-then rules) that would be used to compose chains of inferences within an architecture referred to as a *production system*.[9]

In a paradigm shift in the mid-1980s, attention began to shift from logic-based methods to statistical approaches for handling uncertainties associated with the complexity of real-world problems, such as applications in medical diagnosis and decision support. Representation and reasoning machinery were developed for harnessing probability theory and decision theory,[10] including *Bayesian networks*[11] and, more generally, *probabilistic graphical models*.[12] Systems were developed using these probabilistic representations for making inferences, such as inferring medical diagnoses from information about a patient's illness, sets of symptoms, and lab results. In some systems, the collection of additional information to help refine conclusions or diagnoses was guided by computing the *expected value of information* of additional observations, tests, or data.[13] In addition, AI research scientists began to incorporate and extend techniques developed in the

related disciplines of Operations Research, such as *Markov decision processes* to support sequential decisions.[14]

Despite the rise and fall of excitement in different methods, efforts have continued within and across multiple fundamental representation and reasoning methods. For example, today's successes and focus of attention on large-scale neural networks extends in a recognizable line from the nascent work in the early 1960s on *perceptrons* to the most recent developments with methods and systems based on neural networks. Today, studies of symbolic reasoning methods continue, including on mechanisms for integrating symbolic reasoning with neural models to bolster their abilities to perform logic and more general mathematics.[15]

Machine Learning: Foundation of Today's AI

Machine learning involves algorithms that enable computers to automatically improve their performance at some task through experience. Often that experience takes the form of a large dataset (e.g., in systems that learn to classify which new credit card transactions are likely to be legitimate versus fraudulent) by training on large historical datasets of transactions where the correct classification is known in retrospect. In other cases, training experience may involve active experimentation, as in AI systems that learn to play games by using their evolving current best strategy to play against itself, to collect data on which game moves produce a win. Breakthroughs in AI over the last fifteen years are largely attributable to advances in machine learning. Today, machine learning is viewed as foundational to the field as AI moves into the future.

Beyond the aforementioned early research with perceptrons, today's scientific studies of machine learning extend back to numerous early efforts with learning from data or experience. Such efforts include game-playing systems in chess and checkers and research efforts that laid out surprisingly modern sets of concepts, flows, and architectures for machine learning.[16] For example, research on the Pandemonium system by Oliver Selfridge called out principles of salient feature discovery and the use of multiple levels of representation.[17]

Machine learning research accelerated in the late-1990s. During that time, algorithmic advances, construction of prototypes, and undertaking of empirical studies were catalyzed by the fast-paced rise in computing power and data storage capabilities, along with the explosion in the quantity of online data available for research and development. In the mid-1990s, large amounts of data started to become available via precipitous drops in cost of storage, new data capture technologies, and the massive quantity of content and behavioral data coming with the growth of the web.

A tapestry of machine learning methods has been developed over the last thirty years, many extending methods in traditional statistical analyses to handling datasets with larger numbers of variables and cases and frequently aimed at solving aspirational goals of AI. Enabling advances include methods developed in the late 1980s and early 1990s for directly learning probabilistic graphical models from data[18] and for enhancing the efficiency and capabilities of neural network constructions.[19]

Particularly important to where we are today with the science of AI—and powering the fast-paced progress in research and development—are scientific advances with harnessing multilayered neural networks that came to be referred to as a methodology named *deep learning*. Advances in deep learning have propelled AI to unprecedented levels of capabilities and utility. Innovations, stemming back decades, include the method of *backpropagation* for tuning multilayered neural networks with data[20] and *convolution*,[21] an approach to pooling complex signals into higher-level abstractions.

Discriminative and Generative Models

Machine learning methods can be broadly divided into two main capabilities, *discriminative AI* and *generative models*, each with distinct objectives and application categories. Discriminative models take as input the description of some item and outputs a label, or classification, of the item. For example, in the case of a junk email filter, a discriminative model learns to label each input email as either spam or non-spam by analyzing features derived from the email. These discriminative models directly use

the features of the input data to make predictions or classifications, focusing on the relationship between the input data and its corresponding labels. Discriminative models span classic statistical models of logistic regression, algorithms for learning classifiers from tabular data, and deep learning for diagnosis and classification. Examples of discriminative models include leveraging labeled data drawn from electronic health record systems to predict readmission,[22] sepsis,[23] and the onset of infection[24] in hospitalized patients.

Generative models have been front and center in the recent excitement about AI and its applications. Such models replicate the process by which data is generated. By learning the probability distribution of output features given input features, generative models can create and output new data instances that resemble the training data (in contrast to the labels output by discriminative models). Multiple methods have been used in generative AI, including techniques named *generative adversarial networks* (GANs), *variational autoencoders* (VAEs), *diffusion modeling*, and more recently, *transformers* that yield exciting capabilities of generative AI models. Generative models trained on images are now being used to generate novel imagery, as has become popularized in the DALL-E and Midjourney applications. Beyond images, generative methods are being used in a wide range of applications, including the structure and design of protein sequences and the performing of scientific simulations.

Supervised, Unsupervised, and Self-Supervised Learning

The training procedures by which models are constructed in machine learning can be broadly categorized into supervised, unsupervised, and self-supervised learning. *Supervised learning* relies on labeled datasets. The use of such curated data has been the basis of significant advancements in areas like medical diagnosis, image analysis, and speech recognition. *Unsupervised learning* refers to methods that find patterns in data without explicit labeling. Traditional variants of these methods include clustering and anomaly detection, which have been particularly useful in exploratory data analysis.

Over the last decade, a special form of *unsupervised learning*,[25] named *self-supervised learning*, has become very important. Self-supervision is a simple yet powerful idea that has enabled AI systems to learn from vast unlabeled datasets, such as massive corpora, crawled from across the web. One approach to self-supervision is to generate labels automatically by a "fill in the blanks" process of hiding words in text or other types of tokens in datasets and then trying to predict the hidden information. As an example, a model might predict the next word in a sentence or the next frame in a video sequence based on previous words or frames.

Self-supervised learning represents a significant shift in machine learning, moving away from heavy reliance on human-labeled data. This paradigm is unlocking new possibilities across various fields, enabling models to learn from vast untapped datasets and driving innovation in areas where labeled data is scarce or expensive to obtain.

Inflection Point for AI: Deep Learning

We are now experiencing an inflection in AI with an acceleration in the rate of innovation. The acceleration is largely attributable to advances in research and development with *deep neural networks* (DNNs) over the last decade.

Excitement about the potential of DNNs was sparked by surprising results in speech recognition, natural language processing, and machine vision. In 2009, DNN methods surprised the community with an unexpected reduction in word error recognition rates challenging conversational speech recognition tasks, including one named Switchboard.[26] Progress on the Switchboard benchmark had essentially plateaued for over a decade when progress was made with a DNN approach. Shortly after these gains in speech recognition, another DNN model named AlexNet was developed and demonstrated to perform with surprising capability on an object recognition challenge dataset named ImageNet.[27]

Since that time, research and applications with DNNs have exploded with new challenge problems and applications. Over the last five years, neural models have been used in multiple applications, including scene recognition systems used in semiautonomous driving. In another domain, DNNs
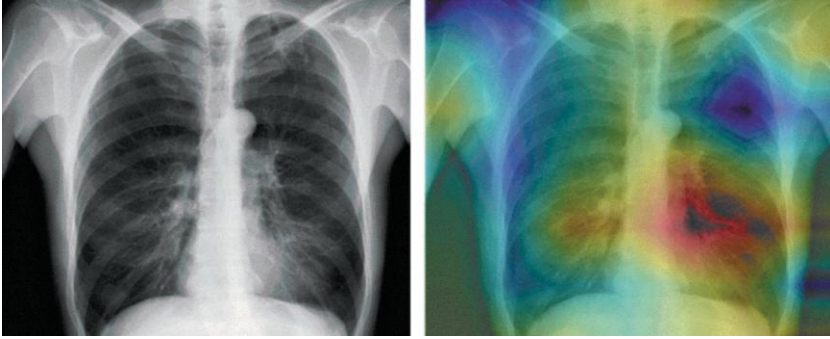
Figure 8.1. Visualization generated by CheXNet model, highlighting a region in a radiological image of the thorax, where the system accurately recognizes pneumonia and highlights the region most representative of the pathology. Pranav Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" *arXiv*, December 25, 2017, https://doi.org/10.48550/arXiv.1711.05225.

have been demonstrated to perform at expert levels with interpreting medical imagery. For example, DNNs have been shown to have the capability to provide expert-level classifications, such as the diagnosis of dermatological disorders from images of skin[28] and diagnoses from radiological films (Figure 8.1).[29]

Sets of evaluation benchmarks have been defined in the language and vision areas, such as the General Language Understanding Evaluation (GLUE), a benchmark formulated to measure the performance of models with language understanding across a range of natural language processing tasks.[30] In stunning advances over a decade, AI systems have reached parity with humans on numerous of the defined challenge problems, as highlighted in Figure 8.2. Details of the progress on the capabilities of AI systems has been captured in the recurrent reports of the AI Index, an annual study of trends in AI hosted at Stanford University.[31]

Key Concepts and Research Directions

Several key directions have come to the fore as important developments, requirements, and directions in work on DNNs, including interest in
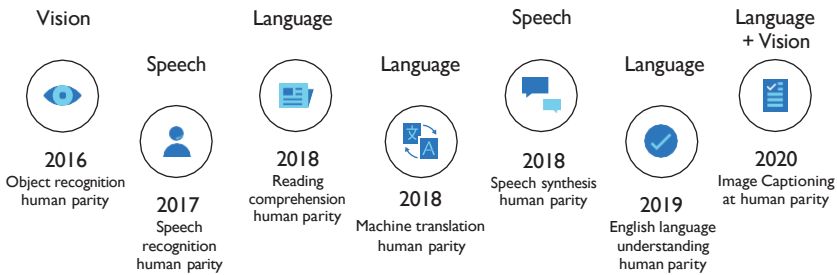
Figure 8.2. AI at an inflection. Deep neural networks have fueled advances in capabilities on benchmarks designed as top challenges for AI systems. This figure shows competencies of AI models where human parity on the challenge problems was reached on seven benchmarks.

automated ways to learn rich representations rather than curate them with expert guidance; efforts on robustness; the weaving together of multi-modal datasets; and the critical value of hardware, innovative algorithms, and programming platforms for research and development.

Learning representations. Early supervised machine learning required the identification of handcrafted salient observations or *features* of the input and based its predictions on those handcrafted features. Researchers have explored how deep learning can identify such features automatically or, more generally, rich representations directly from fine-grained data, in processes referred to as *representation learning*. The automated learning of rich features to represent an image, starting with the lowest-level input pixel features and becoming progressively more complex and abstract at successive layers of neural networks has been a celebrated aspect of modern deep neural models for vision.

The ability to automatically learn, or discover, candidate features enables systems to discover how best to organize the structure of machine learning problems, often yielding more accurate and robust performance on complex tasks than human-defined attributes, albeit at the cost of increased data requirements. Neural models leveraging such representation learning have been developed for natural language processing, computer vision, speech recognition, and health care.

Robustness and generalization. Efforts in the field of DNNs have increasingly focused on achieving robustness and generalization to ensure accurate performance in varying real-world environments that can be assumed to require robust capabilities, for example, accurate diagnoses, classifications, and predictions on new, previously unseen data, that is not contained in training datasets. Efforts in this realm push DNN training procedures to seek universal patterns from their training data so as to reduce their being overfit to training data and to be more adaptable to diverse real-world scenarios.

Hybrid strategies. Successes have been found with combining DNNs with other computational methods such as coupling the neural models with scientific simulations, integrating the methods with Markov decision processes (e.g., reinforcement learning), and integrating DNN approaches with symbolic approaches to reasoning. As an example, the AlphaGo systems rely on an integration of deep neural models for making predictions with reinforcement learning for guiding the choice of actions.[32]

Multimodal modals. Most DNN efforts have focused on the singular modalities of language or visual tasks. In the spirit of pursuing more human-like intelligence, researchers have pursued the development of multimodal models that bring together language, imagery, sounds, and other modalities. Multimodal DNNs include early efforts to do image captioning and more recent efforts to make inferences about language and images for such tasks as writing radiological reports.

Tools, methods, and platforms. With the advent of the importance of DNNs and growing focus of attention on using increasingly large datasets, methods have been pursued for introducing new forms of efficiency via hardware and algorithmic innovation, and for developing programming environments doing exploratory work with architectural designs for neural networks. At the hardware level, graphical processing units (GPUs) have provided speed-ups via parallel processing of matrix and vector operations that are central to deep learning.

Algorithmically, efforts span methods for introducing new forms of speed-ups in distributed computing at the hardware system level as well as on higher-level software innovations aimed at speeding-up the core back-propagation procedure to identify parameters that specify weights of connections in neural models. For example, efforts have focused on adaptation of mathematical optimization procedures like stochastic gradient descent.

Programming environments such as TensorFlow and PyTorch were created to ease the design and testing of DNNs, providing engineers with computing libraries, methods for accelerating GPU computation, and efficiently specifying and revising the structure of neural networks.

Models as platforms. For decades in machine learning, researchers have studied methods for adapting models trained on a source task to perform well on other domains via processes of *fine-tuning* the models with specialized data. This process, often referred to as *transfer learning*, leverages the knowledge that the model has gained from the initial training to perform well on a related, but different, task.[33] Large-scale neural models can serve as platforms for extending via fine-tuning with specialized datasets drawn from target task domains. Given the myriad uses of the large models as platforms that can be extended via domain-specific data, they have been referred to as foundation models.[34] Foundation models can be seen as an extension and scaling-up of transfer learning to DNNs that are trained on extremely large datasets, often encompassing a wide range of topics, languages, or modalities. Their versatility lies in the ability of pretrained models to be fine-tuned with smaller, task-specific datasets, thereby reducing the need for training a model from scratch for each new application. This approach not only saves significant computational resources but also allows for building upon the model's base capabilities and knowledge. The term *foundation* reflects their role as a fundamental base upon which more specialized or *fine-tuned* models can be built, similar to how a foundation supports a structure. Their general-purpose nature and scalability make them akin to a utility or resource that can be tapped into for numerous AI systems. Fine-tuning pretrained foundation models has become a standard

methodology for adding new capabilities, such as adding multimodal capabilities to language-only models[35] and for extending the power of generalist models to specialist performance.[36]

A Second Inflection: Generative AI

The landscape of AI and its influences on the world has now reached a second inflection, *Generative AI*. Generative AI models are rich language and multimodal models that are trained to predict sequences of outputs, given input sequences or *prompts*. These generative models generate the output sequence one item at a time, at each step considering the newest generated item as a new part of the input, as they generate the next item in the sequence. Generative AI spans methods that generate natural language, portions of computer programs, imagery, combinations of imagery and language, and other types of output, such as sequences of amino acids in response to inputs about desired structure and function.

Generative AI systems have been largely based on three innovations that have been brought together to create powerful generative capabilities: the *Transformer architecture*, machinery for self-supervised training on massive diverse content, and a special fine-tuning approach called *instruction tuning*.

Architectural Innovation: "Attention Is All You Need"

A seminal paper introduced the *Transformer architecture*,[37] the foundation of today's generative AI. This particular design of DNN delivers surprising competencies via a mechanism called *attention*, which allows neural language models to *learn to focus on different parts of an input sequence* when generating each part of the output. In short, transformers learn during self-supervised training how to weight the importance of different parts of the input data. Such a broad ability to learn where to look and what to consider has been seen as a pivotal feature for understanding the context and nu-

ances in language, in distinction to earlier approaches for learning about sequences, of only looking adjacently for the context of generation. The power of transformers in various applications, including language translation, text generation, and image processing tasks has led to their broad adoption.

The second pivotal development was combining the Transformer architecture with self-supervised training from a diverse, web-scale dataset. This approach was first demonstrated with the construction of the BERT foundation model.[38] BERT learned language by predicting parts of text that were hidden from it, gaining a broader and more contextual understanding of language via broader learning about where to attend. These innovations laid the groundwork for the development of follow-on Transformer models like the GPT series, LLAMA, and others, each building upon and extending the transformative capabilities introduced by their predecessors.

Alignment with Human Intent and Interaction

A third innovation for enabling modern generative AI is a mechanism for shaping models to follow natural language instructions and to sustain a conversation, versus simply generating tokens that are most likely to follow the input prompt. This process of learning to respond to the intentions of people involves fine-tuning the model on a new dataset composed of various tasks, each linked with explicit instructions and rating the output. The instructions are designed to mimic the way humans would typically instruct each other to perform tasks. The dataset is typically initially generated or refined by human annotators who craft the instructions and provide example outputs or correct responses. To scale instruction tuning, a method referred to as *reinforcement learning from human feedback* (RLHF) is used to expand the instruction dataset and provide measures of the quality of generated outputs. This method involves training and then using an automated approach to scale up the shaping of the model's behavior to ensure a wide coverage of task types and linguistic variations.

Scaling Laws and Emergent Capabilities

A remarkable property of large language models based on the transformer architecture is the existence of a strong empirical relationship between the accuracy of the trained language model and the size of the model (the number of parameters optimized during training), the amount of data on which it is trained, and the amount of computation used during training. This relationship, known as "scaling laws," has been empirically validated multiple times.[39] These scaling laws are important because they predict how larger models trained on larger datasets using greater computing resources yield increased accuracy; if they continue to hold as models are further scaled up, then one can expect even greater accuracy. Figure 8.3 displays a measure of the ability of a learned model to predict next tokens (the "Test Loss"), given a sequence of words at focus of attention, as a function of increases in powers of ten in the compute time, training data, and number of parameters of models.

Scaling laws have provided a reliable framework up to now for predicting basic performance metrics, such as error rates in next word prediction. However, they fall short in anticipating the competencies of models on challenging tasks, including benchmarks in natural language and problem-solving. Training large-scale neural models from broad datasets can be viewed as a form of multitask learning with new tasks being learned with increasing amounts of computation for training and with the size or capacity
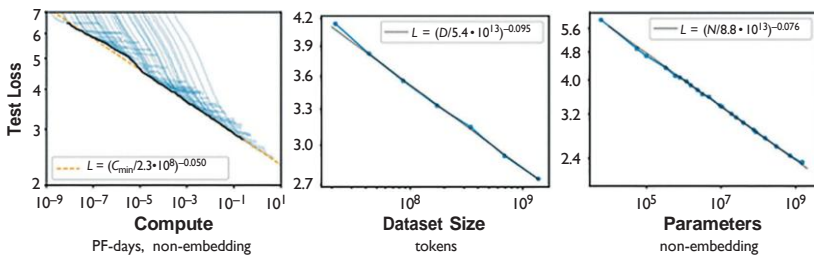

Figure 8.3. Scaling law analyses. Jared Kaplan et al., "Scaling Laws for Neural Language Models," *arXiv*, January 22, 2020, https://doi.org/10.48550/arXiv.2001 .08361.

of models. Task-centric jumps, which have been referred to as the *emergence* of new capabilities, have been observed in neural language models on diverse tasks at different thresholds of model parameters, compute power, and training corpus size. Emergent behaviors include the relatively rapid increase in performance on benchmarks after reaching particular threshold levels of investments in computation for training, as captured in Figure 8.4.[40] Emergent capabilities include jumps in performance on nuanced language understanding benchmarks and with the acquisition of higher-level abilities, such as "theory of mind"—the ability of AI systems to solve challenges with interpreting and predicting the intentions, desires, and beliefs of people.[41] To date, we have a poor understanding of the basis for such jumps in capabilities as a function of model size, extent computation,
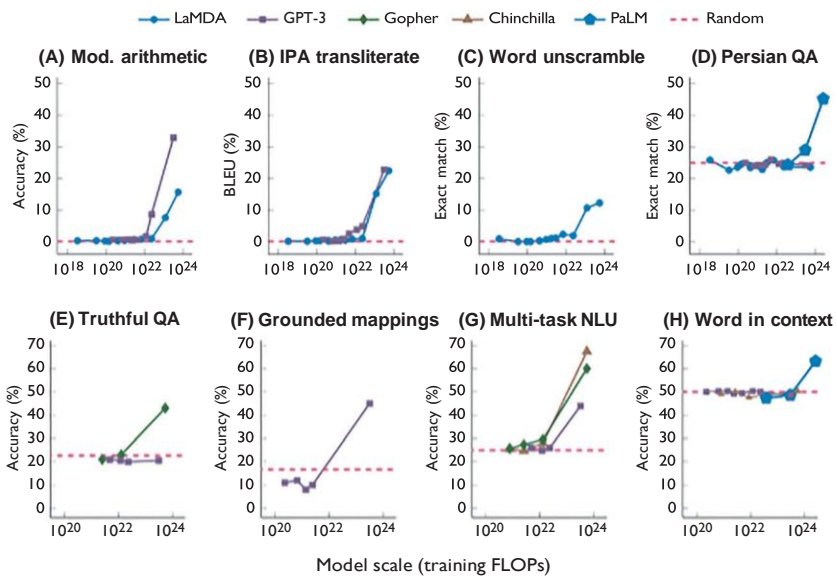
Figure 8.4. Jumps in capabilities on eight reasoning benchmarks for five different generative language models as a function of the number of floating-point operations (FLOPs) invested for optimizing model parameters during training. The jumps have been referred to as the emergence of specific capabilities at particular thresholds of model sophistication. Jason Wei et al., "Emergent Abilities of Large Language Models," *arXiv*, October 26, 2022, https://doi.org/10 .48550/arXiv.2206.07682.

and training data, and links between the accuracy of next word prediction and performance on the more sophisticated benchmarks.

To date, the exact mechanisms and thresholds that trigger emergent capabilities remain largely unpredictable and are an important research direction. This unpredictability underscores a significant frontier in AI research, where the confluence of parameters, computing resources, and training data size creates a complex landscape, within which unexpected and sophisticated AI capabilities can spontaneously manifest.

Surprising Powers of Abstraction, Generalization, and Composition

The original set of ChatGPT systems, using GPT-3.5, GPT-4, and related models have surprised the world with their generalist powers to perform abstraction, generalization, and numerous forms of composition. The models also show broad "polymathic" capabilities, demonstrating the ability to weave together concepts and content drawn from multiple disciplines. The scientific community does not yet have a good understanding of the emergence of their abilities to perform various kinds of summarization, text generation, problem-solving, code program generation, and conversational dialogue. Multiple projects are underway to probe the powers and failings of these models.

Since the release of GPT-4, and related large-scale models such as Claude and Gemini, numerous studies and associated papers have probed potential uses and have provided an array of evaluations. An early survey of capabilities was undertaken by Bubeck and colleagues,[42] spanning a broad set of computing problems, specialist challenges, and the handling of needs and interpretation of events of daily life. The survey highlighted surprising capabilities as well as weaknesses and future directions. Weaknesses include the tendency of the large models to confabulate with the creation of erroneous but persuasive generations and solutions and failures to perform basic arithmetic operations. Studies have also uncovered potentially fundamental challenges with limited abilities to solve complex planning problems that have relied on traditional AI problem-solving on searching through options with backtracking.[43] These challenges have been attributed

GPT-4

Prompt: Can you writea proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces LᴬTEX compiling to following picture.]
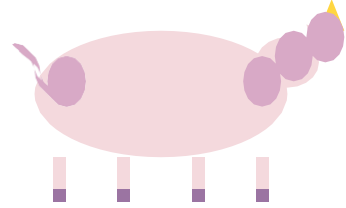
Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8], y1 = [2, 3, 5, 18, 1], y2 = [3, 5, 6, 8, 1], y3 = [5, 1, 2, 3, 4], y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x-axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]
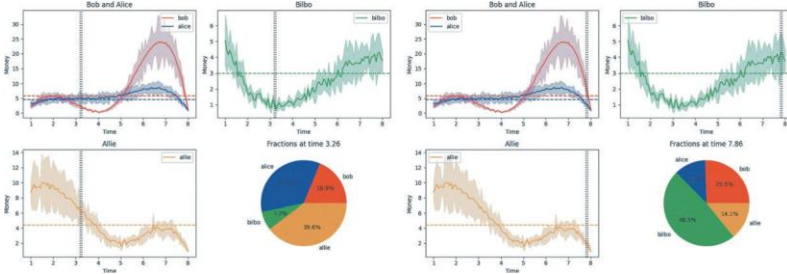
Prompt: Suppose $g(x) = f^{-1}(x), g(0) = 5, g(4) = 7, g(3) = 2, g(7) = 9, g(9) = 6$ what is $f(f(f(6)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g. That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g, and vice versa. Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$. Therefore, $f(f(f(6))) = f(f(9)) = f(7) = 4$.

Figure 8.5. A sample of diverse prompts and output to an early version of GPT-4. Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv*, April 13, 2023, https://doi.org/10.48550/arXiv .2303.12712.

You will be given a name of an object (such as Car, Chair, Elephant) and a letter in the alphabet. Your goal is to first produce a 1-line description of how that object can be combined with the letter in an image (for example, for an elephant and the letter J, the trunk of the elephant can have a J shape, and for the letter A and a house, the house can have an A shape with the upper triangle of the A being the roof). Following the short description, please create SVG code to produce this (in the SVG use shapes like ellipses, triangles etc and polygons but try to defer from using quadratic curves).

Round 1: A pig and the letter H.
Description: The pig can be lying on its back with its four legs forming the two vertical bars of the H, and its curly tail forming the horizontal bar.

Round 2: A house and the letter U.
Description: A house with a U shape, where the two vertical lines of the U are the walls and the horizontal line is the floor. The roof is a triangle above the U.

Round 3: A chair and the letter K.
Description: The back of the chair can have a K shape, with two vertical lines and a diagonal line connecting them. The seat of the chair can be a rectangle.

Round 4: A dog and the letter Q.
Description: The dog is sitting with its tail curled around its body, forming a Q shape. The Q in filled with orange and has a black outline.

Round 5: A car and the letter E.
Description: The car has an E shape on its front bumper, with the horizontal lines of the E being lights and the vertical line being the license plate.

Figure 8.6. Prompts and output demonstrating surprising powers of "compositionality" demonstrated by an early version of GPT-4. Bubeck et al., "Sparks of Artificial General Intelligence."

to the sequential generative processes of current models. Exploratory efforts have pursued insights about the root cause of failures, such as weaknesses models can exhibit with accurately solving constraint satisfaction and mathematics problems.[44]

Tapping Specialist Performance via Steering

For years, specialist performance with large language models has been achieved via training with domain-specific datasets, such as with the construction of BioBert[45] and PubMedBert[46] or fine-tuning foundation
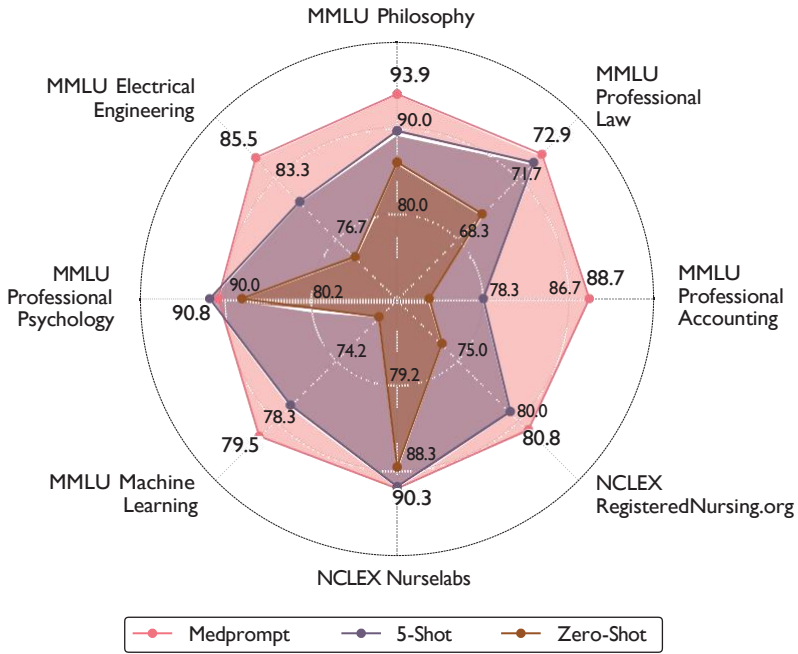
Figure 8.7. Prompting strategies can be used to guide generalist models to act as specialists. This figure shows comparative analysis of simple versus more sophisticated prompting strategies for steering GPT-4 to perform as a specialist on competency benchmarks in multiple realms. Harsha Nori et al., "Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine," *arXiv*, November 27, 2023, https://doi.org/10.48550/arXiv.2311.16452.

models with domain-specific data to update the parameters of the general models via optimization. In addition to surprising powers of abstraction, generalization, and composition, recent studies have demonstrated that generalist foundation models can be guided through special prompting strategies to perform as top specialists. For example, prompting methods can guide GPT-4 to act as a top medical specialist, with record performance on the MedQA benchmark of medical challenge problems.[47] Innovation with prompting shows that generalist models can be steered to perform as experts on competency exams in other areas, including electrical engineering, machine learning, philosophy, accounting, nursing, and psychology.

Research Directions on Generative AI

A great many questions have been framed by the successes and failures of generative AI models. The current questions and curiosity frame a set of research directions and underscore the critical importance of furthering the scientific study of the methods and models.

Representation and reasoning. There is evidence that pushing Transformers via intensive optimization to become increasingly better at predicting the next tokens in their generations, under bounded computing and representational resources, induces the models to induce rich world models as an ideal form of compression. Although several directions have provided insights about the construction of world representations, much remains unknown,[48] and this is an open and interesting area of research.

In a related direction of research focused more on the microstructure of internal activity within transformers, researchers have begun to study the finer details of the activity of the artificial neurons in neural networks that form large language models, as well as the associations among neurons or "neuronal subcircuits" that are induced during training[49] and patterns of neuron activation at inference time.[50] One hypothesis is that a large amount of diverse content forces neural networks to learn generally applicable and special-purpose circuits that can support multiple tasks. Such investigations occur largely in smaller models under controlled learning settings. In such work, small models may be promising as more penetrable, understandable "drosophila," with results that are generalizable to much larger models, just as smaller animal models are used to do medical research aimed at advancing human biology and health care.

Opportunities for more fundamental research include investigations of how principles and methods of probability and decision theory might be more deeply harnessed in representations and inference methods to guide the allocation of computational effort and the selective gathering of information in learning and reasoning.[51] Another direction is to address challenges noted with the ability of generative models to perform planning of the form solved by methods developed in the AI and Operations

Research communities for formulating multistep plans via exploration with search and backtracking.[52] We also see opportunity to move beyond solving single prompts and problems with relatively fixed models to extended presence and situatedness. Directions include exploration of methods aimed at continual reasoning about streams of problems over time.[53] Other opportunities include pursuing understandings and extensions of how the models perform and seeking deeper understandings of challenges and opportunities with the physical embodiment of systems, where grounding of concepts and implications of action are developed with flows of information and learning garnered from immersion in rich, realistic environments.[54]

Memory, learning, and adaptation. Deep neural models do not have the ability to quickly learn and adapt as humans do to real-time experiences and information. Once they are trained, these models are then applied but typically remain fixed, or sometimes they are updated via the traditionally long cycle times of fine-tuning. Long cycles for collecting data and building updated models means that late-breaking scientific advances, news, and information will be unavailable to large language models without the use of special machinery to augment inferences. Efforts to address these challenges include extending large models with methods for search and retrieval of recent information. While these adjuvant techniques are helpful, new methods and machinery that enable faster-paced and near real-time memory and learning would be game changing. Opportunities include developing and integrating methods for ongoing, never-ending learning.[55] Extending abilities to remember, learn, and adapt would enable models to stay up-to-date and would enable breakthroughs in personalization.

Architectural innovation. The Transformer has been a go-to architecture for generative AI. Nonetheless, this architecture and methodology has limitations, such as challenges with handling long-term dependencies in sequences. There are opportunities to innovate with new architectures, including introducing new mechanisms into Transformers.

Reliability, calibration, and trustworthiness. As AI systems become more integrated into daily life, ensuring their reliability and safety is paramount, especially when the methods are applied in high-stakes areas like medicine, criminal justice, education, and industrial process control. Characterizing and communicating potential errors, including erroneous generations and rates of false positives and false negatives in pattern recognition, is critically important in understanding costs of failures. Considerations of types of failures and their rates of occurrence is important in ethical deliberations about uses of AI in specific domains and contexts; AI capabilities and errors frame cost–benefit considerations and decisions that hinge on value considerations.

A weakness of generative AI models is their propensity to generate content that is persuasive yet erroneous. A critical research direction is to develop methods and machinery for assigning well-calibrated confidences to generations and also to deepen understanding of when hallucinating content is expected and desired (e.g., generating fiction) or is a concern (e.g., performing medical diagnoses). Directions include developing internal machinery, fine-tuning, experimenting with new forms of prompting, and calling external tools, such as databases and search engines that perform traditional information retrieval for providing verification and constraints. Recent work has explored careful curation of high-quality datasets, including using large-scale models to generate high-quality data to boost the efficiency of learning and accuracy of inferences.[56]

Some studies have verified good calibration of confidences in specific settings. For example, Figure 8.8 shows good calibration of the confidence of GPT-4 about its answers to multiple choice challenges on competency exams in medicine.

Power of small models. While scaling laws, confirmed by empirical studies and theoretical results,[57] suggest that large scale is need for top performance, recent work has demonstrated remarkable power with smaller models, some built from high-quality datasets. In recent work, large language models are used to supply training data to build more compact models that show strong performance.[58] Research is needed to better understand how one can achieve strong capabilities with smaller datasets and computational
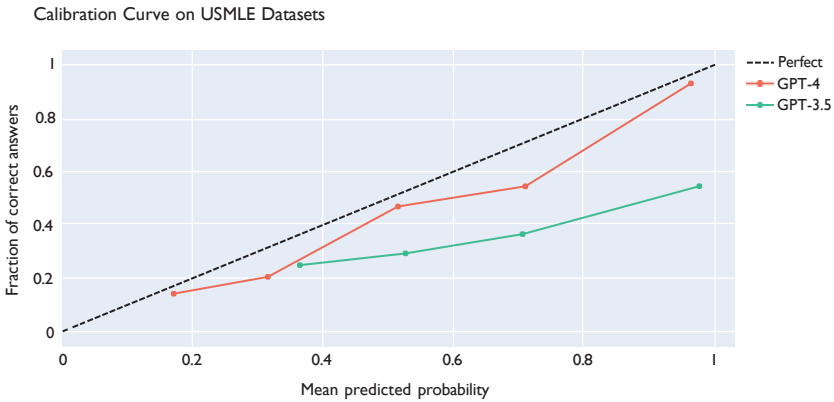
Calibration Curve on USMLE Datasets

Figure 8.8. Calibration of confidence of GPT-4's answers in response to challenge problems drawn from medical competency exams. Harsha Nori et al., "Capabilities of GPT-4 on Medical Challenge Problems," *arXiv*, April 12, 2023, https://doi.org/10.48550/arXiv.2303.13375.

resources, including questions about whether such model construction depends in some way on the poorly understood special properties of data generated by the larger models.

Grappling with opacity and complexity. Large-scale neural models are difficult to understand, potentially hindering scientific progress dependent on insights about the induction of neural circuits and larger representations. New tools are needed to better understand representations and inference in large-scale models. There is a growing similarity of the "black box" challenges of large-scale neural models and the difficulties of probing the operation of biological nervous systems with fine-grained unit recordings and coarse-grained fMRI and related studies. There is a potential convergence of methods and analytical tools between these fields.

Mechanisms and designs for human–AI collaboration. There are great opportunities to extend prior work on human–AI collaboration.[59] Although generative AI models are trained to engage in conversational dialogue, there is a large space of possibilities to design interaction strategies that emphasize the primacy of human agency in problem-solving and that introduce

new styles of human–AI interaction that enable AI systems to complement human decision-making.[60] Related goals include developing better ways for generative AI systems to share the rationale for their generations and recommendations.

Engineering Trends with Generative AI

Although it is impossible to predict the future, especially in an area as dynamic as generative AI, it is nevertheless useful to examine emerging trends in the technology that may shape the future of the technology and its applications. In this section we consider several trends that have emerged after the November 2022 introduction of ChatGPT and their potential to change the future of generative AI.

Multimodal generative AI models. Whereas initial LLM models were trained only on text data, one recent trend is toward training models on multimodal data, such as text, image, video, and sound data. For example, Google recently released a model that it calls Gemini, trained "from the bottom up" on such multimodal data. In October 2023, OpenAI made available a version of its GPT-4 system, GPT-4V, which can accept image and text data as input (although its output is still text only). Figure 8.9 shows a typical interaction with GPT-4V, in which it is able to interpret the content of an uploaded image and reason about how to stack the items in a stable fashion.

The significance of this trend toward multimodal models is that such models hold the potential to capture significantly more commonsense knowledge about the physical world—knowledge that cannot be easily captured in text alone. If successful, this trend could lead to significant new applications, for example, systems that observe and guide people step-by-step as they cook a particular recipe for dinner or as you assemble a new piece of furniture. One interesting question is whether successful development of such multimodal models might cause a rapid burst of new progress in robotics, given that much of what limits robotic systems today is their poor ability to interpret and reason about physics of diverse objects and environments.

How can I stack these four objects in a stable vertical stack?

To stack the four objects in a stable vertical stack, consider the following based on their apparent size, shape, and weight:

1. **Base:** Start with the largest and heaviest object as the base to provide maximum stability. Here, the black rectangular object (which appears to be a case or pouch of some kind) seems to be the best choice for the base given its flat, wide shape.

2. **Second Layer:** Place the coffee mug upside down on top of the base object. The flat bottom of the mug should provide a stable surface for stacking. The handle might pose a bit of an issue for balance, so ensure it's positioned in a way that doesn't cause instability.

3. **Third Layer:** The coin seems to be the next logical choice. Place it flat on the bottom of the overturned mug. Its slim profile and weight should allow it to rest securely.

4. **Top Layer:** Finally, place the green object, which appears to be some type of plastic tool or toy, on top of the coin. Its lightweight nature and tapered design should allow it to balance on the coin, but be cautious as this might be the least stable part of the stack.

Ensure that each object is centered as much as possible on the object below it to maximize stability. Also, be aware that this stack, particularly with the smaller objects on top, might be inherently unstable and can be knocked over with slight disturbances.

Image input to GPT-4V

Result of following instructions from GPT-4V

Figure 8.9. An interaction with GPT-4V, which accepts image as well as text inputs. Here the input image on the left shows four items on a desktop. The input request to GPT-4V is "How can I stack these four objects in a stable vertical stack?" When the output answer from GPT-4V (shown in the middle) is followed, it produces the vertical stack shown on the right. Created by Tom Mitchell using GPT-V website.

Power of synthetic data. Generative models and more traditional simulation methods are being used to generate large quantities of training data that are being used successfully to build and extend neural models. Datasets being generated and harnessed includes visual datasets and focused, high-quality distillations of specific types of output, such as reasoning strategies[61] and domain-specific data.[62]

Incorporating software plugins. LLM's like GPT-4 exhibit many impressive abilities, they also have many limitations and shortcomings. For example, today's LLM's cannot reliably perform arithmetic with large numbers (e.g., multiply 483 times 9,328) and can hallucinate incorrect answers to factual questions. Model plugins consist of traditional software (e.g., a calculator, a database of factual information) that can be called as subroutines by LLMs. Providing LLMs with plugins allows them to overcome numerous limitations and to take advantage of the vast store of software developed by many groups over multiple decades of effort. For example, as of November 2023, ChatGPT had access to approximately 1,000 plugins—from calculators, to web search engines, to restaurant reservation apps—which

MMLU Philosophy
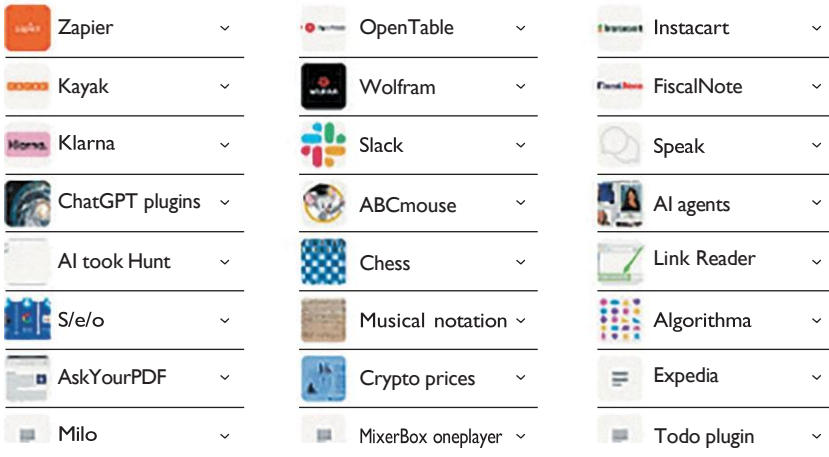From sources across the web

Zapier ⌄	OpenTable ⌄	Instacart ⌄
Kayak ⌄	Wolfram ⌄	FiscalNote ⌄
Klarna ⌄	Slack ⌄	Speak ⌄
ChatGPT plugins ⌄	ABCmouse ⌄	AI agents ⌄
AI took Hunt ⌄	Chess ⌄	Link Reader ⌄
S/e/o ⌄	Musical notation ⌄	Algorithma ⌄
AskYourPDF ⌄	Crypto prices ⌄	Expedia ⌄
Milo ⌄	MixerBox oneplayer ⌄	Todo plugin ⌄

Figure 8.10. Small sample from approximately a thousand plugins accessible to GPT. From the OpenAI website.

significantly extend its capabilities beyond those provided by its trained neural network (Figure 8.10).

The model decides whether and when to invoke any given plugin, depending on the prompt it is responding to, but at present most generative AI models limit the number of plugins to be considered in any given conversation. For example, ChatGPT requires users to preselect at most a handful of its available plugins for any given conversation. It remains to be seen how large a set of plugins a model will be able to automatically consider invoking. However, giving models access to the vast store of software developed across the computer industry will be a goal for future systems. One question raised by the rise of plugins is whether generative AI models will become user interfaces of choice to many software packages that currently have their own idiosyncratic interfaces. Will future users prefer to interact in natural language conversation instead of learning the specialized interface for each software application? Plugins are extensions that allow LLMs to impact the world beyond conversations, such as researching.

Beyond such tasks as arithmetic calculation and information retrieval, plugins can enable LLMs to perform myriad functions, including executing

actions in the open world, such as making purchases, sending messages, and controlling physical systems. While such integration with broader software and systems can provide new functionalities and services, the new powers also pose risks to safety and security and must be handled with care.

Multifunctional interactive workspaces. In a direction of innovation related to plugins, we see the rise of integrated interactive experiences that promote human–AI collaboration by enabling users and AI components to work together in a step-by-step manner on problem-solving, where multiple tools, data analysis abilities, and code creation are made available in a collaborative approach to solving problems and subproblems. A portion of a sample session of such a multifunctional interactive workspace provided by OpenAI named Advanced Data Analysis is displayed in Figure 8.11.
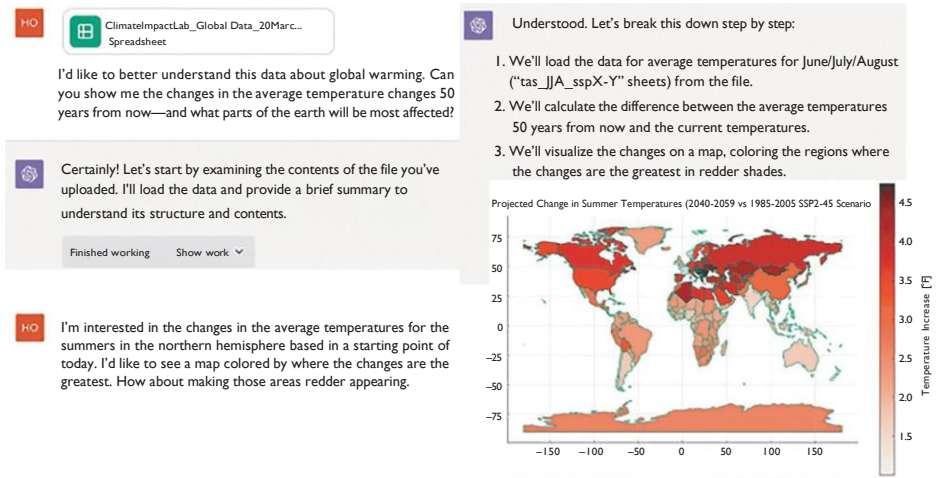
Figure 8.11. Advanced Data Analysis provided by OpenAI, a multifunctional interaction workspace that enables databases and papers to be loaded for analysis, and provides multiple step analyses, introducing tools as needed, including writing of code and provision of visualizations, and with ongoing sharing of plans and steps with users. Created by Eric Horvitz using OpenAI Advanced Data Analysis, November 2023.

Software development environments for programming with AI models. In contrast to using tools that enable generative AI models to call other software as subroutines, this trend supports the development of new software systems that instead call generative AI as subroutines. Frameworks such as the open-source LangChain, Microsoft's Semantic Kernel, and AutoGen have emerged to support software developers in building systems that call multiple instances of a generative model.[63] These frameworks make it easier to build software systems that capture the benefits of LLMs (e.g., to interact in natural language, and to perform certain types of commonsense reasoning) while also incorporating standard programming and capabilities missing from generative AI, such as long-term memory and database access. One aspect of generative AI that makes this especially interesting is the ease with which one can "program" or "instruct" an instance of a generative AI model on how to behave. For example, Figure 8.12 shows the text used to instruct, or program, an instance of GPT-4 to perform the role of

1. Instructions to GPT-4 to define StudentBot (Rufle):

You are an enthusiastic 18-year-old student who is trying to learn. You need the user (who is a teacher) to slowly teach you all topics in the material. You have access to a list of topics, but not to the material itself. You must learn one topic at a time. This is the list of topics you found from the internet that you need the user to slowly teach you (by order): + {tutoring-script}

You need to learn very little at a time. Ask the user (who is the teacher) to teach you the material, little by little. If the teacher gives the answer, you must (a) show appreciation and understanding; (b) insert [SMILE]; and then (c) ask a follow up question if you need more information about the current topic or ask a question about the next topic. Do not move on the next question before getting an answer for the current question. If the teacher doesn't know something, tell the teacher you will be thrilled if the teacher can check it and get back to you. When all the topics are covered, thank the teacher, say I've asked all the questions.

Figure 8.12. Natural language instructions used to "program" an instance of GPT-4 to play the role of an artificial student, as part of a larger online educational software, in which humans learn by teaching this artificial studentbot (implemented by GPT-4), with the occasional assistance of an artificial ProfessorBot (implemented by a second instance of GPT-4). Robin Schmucker et al., "NeurIPS Paper 38: Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems," 2023, https://neurips.cc/virtual/2023/79097.

an artificial student, to be taught by a human teacher within an online education application. The programming of the LLM is done here using only natural language instruction rather than a programming language.

One interesting question about the future which is raised by this trend is whether we are beginning to see the emergence of a new paradigm for software development which, unlike previous paradigms that relied exclusively on formal programming languages to instruct the machine, will in the future seamlessly blend natural language instructions with formal languages (Figure 8.13).

Personalized generative AI systems. Generative AI models such as OpenAI's GPT-4 and Google's Gemini are very costly to develop and are so large (containing hundreds of billions of learned parameters) that they are not downloaded, but only used remotely over the web. As a result, it may seem unlikely that these models could be personalized to each of billions of people on the planet. Nevertheless, we are already beginning to see a trend toward personalized LLMs. For example, ChatGPT allows users to provide a natural language description of themselves and their interests which it can use to
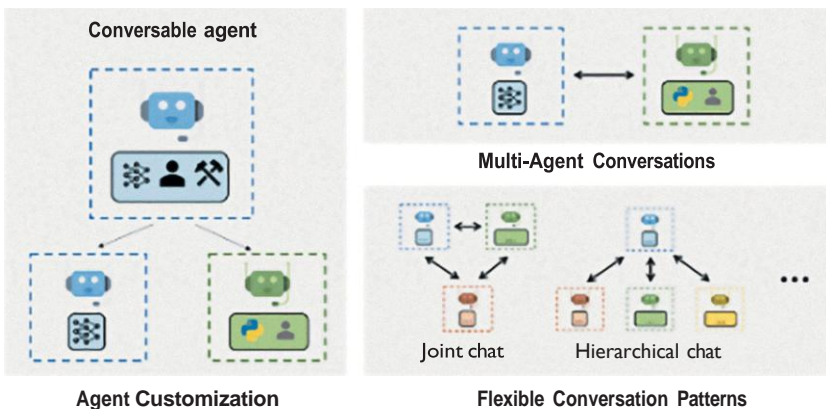
Figure 8.13. AutoGen orchestration framework for generative AI models allows the efficient specification of roles and flows of generations. Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," paper presented at the Conference on Language Modeling, Philadelphia, PA, October 7–9, 2024.

modulate conversations with that user (e.g., to customize to their educational background). Furthermore, Google has released an experimental version of its conversational assistant Bard that enables users to give it access to their entire Gmail collection as well as their online Google Docs, then discuss the content of these. For example, Figure 8.14 shows a typical interaction with this experimental version of Bard. Beyond this, Microsoft has also released a new version of their Office software suite in which LLMs are integrated with systems such as Word, PowerPoint, and Excel. Both Apple and Google have announced plans to release versions of LLMs small enough to run on their respective mobile phones, opening the possibility of highly personalized LLM-based agents that preserve privacy by operating solely on personal devices.
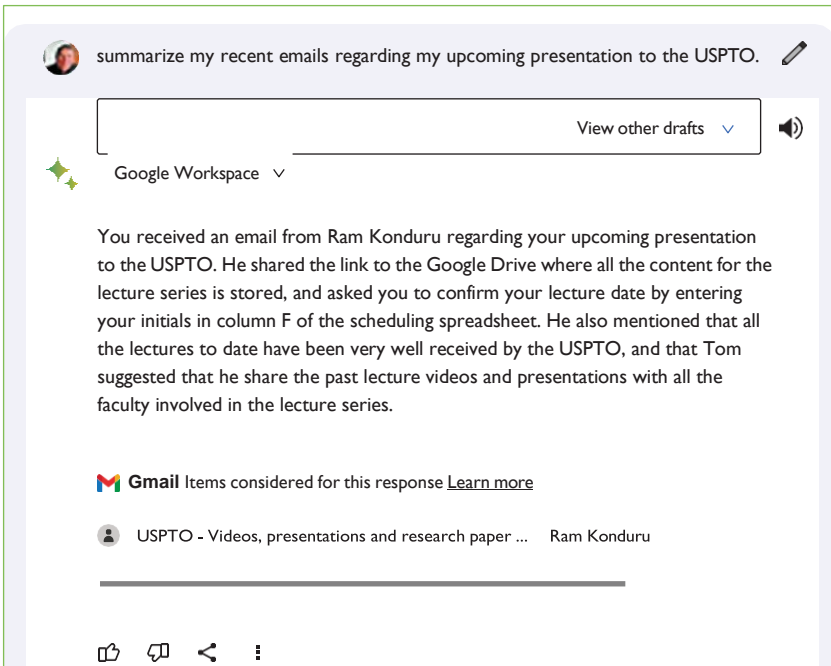
Figure 8.14. A conversation with Google's Bard about the content of the user's Gmail and Google Docs. Created by Tom Mitchell using the Google Bard web interface.

The significance of this trend is that it suggests that the future will see an increasing ability of generative AI systems to interface with personal data, and data of corporations, in ways that make them tremendously more useful and knowledgeable about the problems of interest to their users. Customization to specific users, corporations, and problem settings is likely to be supported by a combination of model fine-tuning, providing access to relevant user data, and direct natural language instructions defining roles for the agent.

Open-source models. One trend in generative AI might be summarized as "bigger is better." Between 2018 and 2023 the sequence of top state-of-the-art generative AI models followed a clear scaling law: models with more parameters, trained on larger datasets, produced significantly improved capabilities (Figure 8.15). This led to models with costs of over $100 million to train and containing so many parameters that they would not fit on most computers. Given this trend, one might expect a future in which only a few dozen well-resourced companies and governments could afford to develop the next generation of models, and where the rest of us would only be able to access those models over the cloud. As mentioned earlier in the discussion of research directions, a number of new models being developed and fielded rely on many fewer parameters—few enough that the models can be downloaded and trained or fine-tuned on much smaller computers. Although these smaller models do not match the competence of the very best models, they exhibit surprisingly good competence, especially when trained for specific domains such as medicine or finance, and when trained using carefully selected training data such as textbooks. These small models make it feasible for researchers and developers across the world to build and work with generative AI, rather than just the employees of a handful of organizations; that is, they make open-source shared development by many cooperating developers possible.

Of all the trends mentioned here, this trend toward smaller, open-source, widely shared models may be the most consequential, as it will strongly influence both the number of researchers and developers who participate in advancing the technology, and it will strongly influence the ability of

So far, bigger models trained on larger data sets produce best results

But models orders of magnitude smaller and cheaper are surprisingly good

Model	Parameters	Year
Falcon	40B	2023
Alpaca	7B	2023
Vicuna	13B	2023
Phi 1.5	1.3B	2023
Phi 1 small	350M	2023

Model	Parameters	Year
GPT-1	117M	2018
BERT	345M	2018
GPT-2	1.5B	2019
GPT-3	175B	2020
PaLM	540B	2022
Megatron-Turing	540B	2022
GPT-4	??	2023

Figure 8.15. Sizes and year of release of various generative AI models. Created by Tom Mitchell.

governments to control and regulate uses of the technology and the "guardrails" placed on it.

Consider first the impact of the open-source trend on the number of technical experts who can work to advance the technology. Because current state-of-the-art models such as OpenAI's GPT-4 and Google's Gemini are so large and so expensive to train, they can only be accessed over the cloud, and the next generation of these models can only be developed by organizations such as OpenAI, Google, Microsoft, Amazon, and other organizations who have computational infrastructures that cost hundreds of millions of dollars. Such organizations may have many thousands of employees, but this number is dwarfed by the number of researchers and developers outside such large organizations (e.g., university faculty and students in computer science, and employees at small startup companies). Because the rate of research progress is often strongly dependent on the number of researchers working on a problem, a successful and vibrant open-source movement is likely to result in more rapid advances and in the democratization of application development. One concern of the US government as it seeks policies that enable the United States to lead in this technology is the potential loss of university research as a key driver of AI advances. For many decades, US universities drove the key advances in AI. However, in recent years the greatest AI breakthroughs have instead come from industry, because universities lack the high-cost computational

resources necessary to train and experiment with the largest, most advanced AI foundational models. One proposal under consideration is to fund a National AI Research Resource (NAIRR) to provide computational resources to keep US universities a vital part of research at the frontier of AI. A pilot NAIRR effort is being organized by the National Science Foundation and is planned for launch in mid-2024.

The success or failure of smaller models and therefore of the related open-source effort in generative AI will also have a strong impact on whether and how governments can track and regulate AI technology. Large corporations that work in this area are already cooperating with various governments to create frameworks, best practices, and regulations to minimize the risk of AI being used for nefarious purposes, as well as risks of adverse unintended consequences. If only very large AI systems dominate in the future, then the open-source movement is likely to be small or nonexistent, and governments can continue to work with large corporations and can effectively enforce any government regulations. However, if small AI models and the corresponding open-source movement succeed, then it will be very difficult, perhaps impossible, for governments to know which organizations and which individuals have highly capable AI models and what they are using them for. In short, if small AI models become highly capable and easily copied and ported, then they will become very difficult to regulate.

Key Opportunities with Applications

Discriminative and generative AI models have great applications in daily life and in specific domains and specialties. Major areas of future impact include the biological and physical sciences, health and well-being, and education.

Biosciences. AI's impact is expanding rapidly in the biosciences. AI methods promise to provide fast-paced leaps in understanding complex biological processes and designing new drugs and therapies. Neural modeling pipelines, including AlphaFold[64] and RoseTTAfold,[65] are providing game-changing capabilities to biologists. Recent work on harnessing these and

other neural modeling methods are putting tools in the hands of biologists for estimating protein structure and better understanding protein function and interactions. As an example, AI tools were recently used to perform a cross-proteome, large-scale screening of potential protein–protein interactions in yeast cells (Figure 8.16). The screening identified previously unknown protein interactions in these eukaryotic cells—cells that are closely related to those that we are composed of.[66] Many of the interactions could be mapped to pathways by biologists. However, the roles of several predicted interactions remain mysteries, framing new questions in cell biology. Advances with predicting protein–protein interactions offer a multitude of possibilities for harnessing AI advances for understanding and intervening with cellular pathways. Figure 8.17 shows that recently developed *diffusion modeling* techniques, analogous to AI methods for image generation, have been harnessed in protein design.[67] Such methods can be harnessed for designing new medications, protective binders that block the active site of viruses, and synthetic vaccines. Over the next decade and beyond, AI could revolutionize personalized medicine, offering tailored treatments based on illness specifics and individual genetic profiles, and accelerate the pace of biotechnological innovation, possibly leading to solutions for today's incurable diseases.

Health care. To date, AI has been a sleeping giant in health care. In the next decade, we may see AI becoming a regular assistant in diagnosis and treatment planning, offering more accurate and faster diagnoses. AI could also enhance remote health care and monitoring, making quality health care accessible in underserved regions. Multiple opportunities for traditional machine learning exist, as do uses of discriminative and generative neural models to assist with diagnoses and predicting outcomes. Work to date has demonstrated great possibilities for enhancing the quality of care, including raising levels of diagnostic and therapeutic excellence, and reducing human errors. Beyond clinical decision support, the capabilities of generative models to generate and summarize reports can reduce the administrative on physicians providing them with more time for quality patient engagement (Figures 8.18 and 8.19).
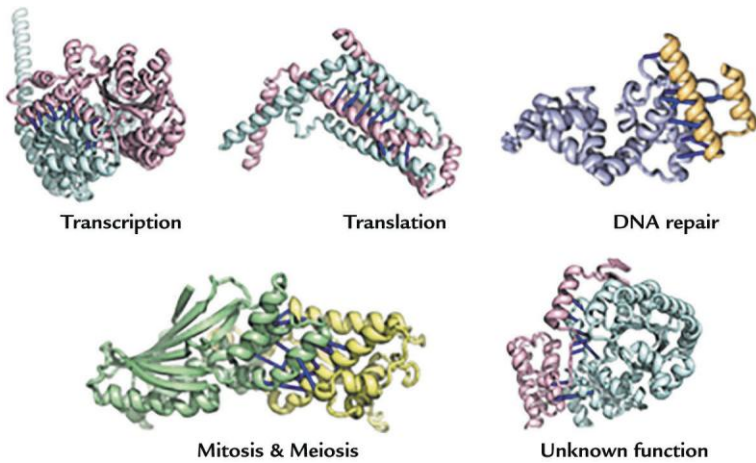
Figure 8.16. Decoding protein complexes. In work moving beyond structure, DNNs have been applied to identify likely protein complexes in eukaryotic cells. The complexes have been linked to processes of transcription, translation, DNA repair, mitosis and meiosis, metabolism, and protein transport within cells and across membranes. The dark blue lines indicate likely points of contact predicted between the proteins. The function of some of the identified complexes are mysteries. See Ian R Humphreys et al., "Computed Structures of Core Eukaryotic Protein Complexes," *Science* 374, no. 6573 (December 10, 2021): eabm4805–eabm4805, https://doi.org/10.1126/science.abm4805.
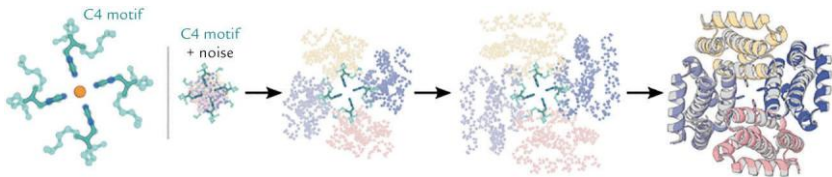
Figure 8.17. Supercharging protein design. Use of diffusion modeling methods to design proteins. In this case, a protein is designed with conditioning on a given motif. See Joseph L Watson et al., "De Novo Design of Protein Structure and Function with RFdiffusion," *Nature* 620, no. 7976 (August 2023): 1089–1100, https://doi.org/10.1038/s41586-023-06415-8.

If you survey the literature of what we know, what new studies might be done to make a contribution forward in resolving the uncertainties?

Given the current understanding of the pathophysiology of glaucoma, potential areas of research to advance knowledge and improve outcomes for patients with glaucoma could include the following:

1. Elevated intraocular pressure (IOP): Further studies could investigate the mechanisms by which elevated IOP leads to optic nerve damage, and identify more accurate and personalized methods for predicting which patients with elevated IOP are at the highest risk for developing glaucoma.

2. Vascular dysregulation: Research could focus on gaining a better understanding of the role of vascular dysregulation in glaucoma, and identifying new interventions to prevent or mitigate optic nerve ischemia and damage.

3. Neurodegeneration: More studies could be done to elucidate the mechanisms involved in neurodegeneration in glaucoma, and to identify new neuroprotective interventions for the prevention and treatment of glaucoma.

4. Genetic factors: Further research could aim to identify new genetic risk factors for glaucoma, and explore how genetic testing and personalized medicine could be used to improve the detection, prevention, and treatment of glaucoma.

5. Clinical trials: Large-scale clinical trials could be conducted to compare the effectiveness of different interventions for glaucoma, including medications, surgeries, and new treatments such as gene therapy and stem cell therapy.

By pursuing these areas of research, it may be possible to advance understanding of the pathophysiology of glaucoma, identify new targets for therapies, and improve outcomes for patients with glaucoma.

Figure 8.18. Snippet from a medical session with GPT-4, showing diagnosis, summarization, review of relevant frontier research, and patient communication. Full session available at https://unlocked.microsoft.com/ai-anthology/case -studies. Created by Eric Horvitz with GPT-4, January 2023; from Eric Horvitz, "Case Studies: Sample Sessions with GPT-4 Prior to Public Release," *AI Anthology, Microsoft Unlocked* (2023), https://unlocked.microsoft.com/ai -anthology/case-studies/.
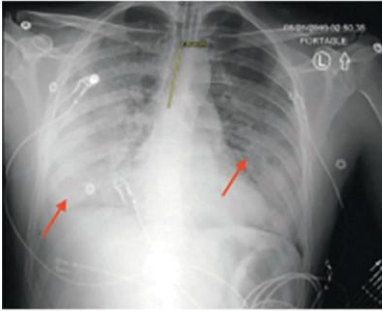
What is shown in this image?

LLaVA-Med: The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.

Are there any devices implanted in this patient?

LLaVA-Med: Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.

Figure 8.19. Rise of multimodal models. Creating a multimodal medical imaging model by fine-tuning a generalist open-source model with millions of aligned images and captions accessed from the openly available medical literature. Chunyuan Li et al., "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (Red Hook, NY: Curran Associates, 2024), 28541–28564.

Physical sciences. Work is underway on numerous fronts in the physical sciences with uses of generative AI models. A detailed review of recent efforts and directions is provided in *The Impact of Large Language Models on Scientific Discovery*, an AI survey by Microsoft Research AI4Science and Microsoft Azure Quantum.[68] In material science, AI is already accelerating the discovery of new materials and understanding complex physical phenomena. Work includes using neural models to provide candidate chemical compounds and to speed up analyses of suitability of candidates by providing efficient approximations of more complex traditional quantum computations. With recent advances in AI-driven simulations and predictive modeling, the next decade could see AI systems designing materials with tailored properties for specific applications, such as ultra-strong composites for aerospace or highly efficient conductors for electronics. Directions with applications of AI for science include the development of large integrated scientific foundation models that form datasets drawn from multiple scientific domains and at a variety of spatial scales.

Climate and sustainability. AI methods are showing promise with optimization of renewable energy systems and with important tasks as predicting

climate patterns and responses to alternate interventions. Looking forward, AI could be instrumental in helping with the discovery and design of more efficient catalysts and overall processes for carbon capture and storage. AI-driven models could offer more precise predictions of climate change impacts, aiding in more effective policymaking and environmental protection measures (Figure 8.20).

Education. GPT-4 is being explored in early deployments, including by Khan Academy, but also in educational research. There are great opportunities to harness generative AI systems to act as a personalized tutor, per the "theory of mind," pedagogical skills, and explanatory capabilities demonstrated by the largest models (Figure 8.21).

Engineering brainstorming and design. Generative AI, including problem-solving, guidance, and visualization of novel designs might provide a transformative toolkit to boost engineers' creativity and innovation. Generative AI models trained to have language, imagery, and multimodal capabilities can help scientists to formulate, explore, and visualize complex concepts or designs that they might not have considered otherwise. There are opportunities for such models to serve as collaborative partners, providing instant feedback or making suggestions based on prompts describing goals. Figure 8.22 shows an early exploration with the DALL-E2 system of visualizations of designs for a combination of solar water heating and power generation.

AI, People, and Society: From Technical to Sociotechnical

The capabilities of AI methods are *dual use*. AI methods can be harnessed in sciences, engineering, and in daily life to raise the quality of life and to promote human flourishing. They can also be leveraged by malevolent actors to pursue costly and criminal activities. Beyond explicit pursuits of uses of AI in adversarial ways, uses of AI may have inadvertent influences on people and society. The intersection of AI with societal aspects encompasses

Wetland emission and atmospheric sink changes explain methane growth in 2020

EH Can you compare the strengths and weaknesses with doing what the authors did with their atmosphere 3D inversion analysis versus an alternate study with using a Bayesian hierarchical model (BHM) for the atmospheric 3D inversion component of the study?

There are different approaches to atmospheric inversion, which is the process of estimating sources and sinks of a trace gas (such as methane) from its observed concentrations in the atmosphere. One approach is the variational Bayesian inversion, which is based on minimizing a cost function that measures the discrepancy between the prior estimates of the sources and sinks and the posterior estimates that are consistent with the observations, weighted by their respective uncertainties. This approach requires an adjoint model that computes the gradient of the cost function and an iterative algorithm that finds the optimal solution. Another approach is the Bayesian hierarchical model (BHM), which is based on specifying a probabilistic model that relates the sources and sinks to the observations, accounting for the uncertainties in the model parameters, the observation errors, and the prior information. This approach requires a Markov chain Monte Carlo (MCMC) algorithm that samples from the posterior distribution of the sources and sinks given the observations.

Both approaches have strengths and weaknesses, and the choice of one over the other may depend on the specific problem, the available data, the computational resources, and the scientific objectives. Here are some possible comparisons between the two approaches:

– The variational Bayesian inversion may be more computationally efficient than the BHM, especially for high-dimensional problems, as it does not require repeated runs of the forward model or the storage of a large number of samples. However, the variational Bayesian inversion may also be more sensitive to the choice of the initial guess, the convergence criteria, and the regularization parameters, and it may not fully explore the

Figure 8.20. Snippet from a session showing analysis of late-breaking scientific paper, showing rich dialog on the scientific methods, foundations, and future directions. Full session available at https://unlocked.microsoft.com/ai-anthology/case-studies/. Created by Eric Horvitz with GPT-4, December 2022; from Eric Horvitz, "Case Studies."

EH The computer is trying to teach a high school student about quantum computing, using analogies and trying to teach the basics so they have a true understanding of the main concepts. The student already studied the basics of quantum physics but is puzzled about quantum computing.

I'd really like to better understand how quantum computers compute and the types of problems they can solve.

Quantum computing is a way of using the properties of quantum physics to perform calculations that are impossible or very hard for classical computers. Quantum physics is the study of the behavior of the smallest particles and waves that make up matter and energy, such as electrons, photons, and atoms.

EH Yes, but how exactly do they compute?

Quantum computers use quantum bits, or qubits, as their basic units of information. Unlike classical bits, which can only be either 0 or 1, qubits can be in a superposition of both 0 and 1 at the same time. This means that a qubit can store and process more information than a classical bit, and that multiple qubits can be entangled, or linked, so that their states depend on each other.

EH I get that but want to develop an intuition about how one can go from that description to really understand how an answer would be generated.

One way to think about quantum computing is to imagine that you have a very large book with many pages, each containing a possible answer to a problem. A classical

Figure 8.21. Snippet from education session on quantum computing, showing rich dialog, signs of pedagogical competence, and responsiveness. Full session available at https://unlocked.microsoft.com/ai-anthology/case-studies/. Created by Eric Horvitz with GPT-4, September 2022; from Eric Horvitz, "Case Studies."

reliability and safety issues, privacy and security trade-offs, and fairness and accountability.[69] Legal and ethical issues around data provenance, intellectual property, and copyright are increasingly pertinent. AI's role in military applications brings up concerns about competitive landscapes and the potential for destabilizing influences. Socially, AI risks exacerbating

Figure 8.22. The powers of composition demonstrated by the multimodal DALL-E2 system provide a glimmer into the potential uses of generative AI as a design colleague. Created by Eric Horvitz using DALLE-2, April 2022.

the digital divide, impacting job markets, and enabling malevolent uses like deepfakes and online manipulation. The deeper social, cultural, and psychological dimensions—trust, authenticity, diversity, agency, and creativity—are also crucial areas for consideration. A great deal of discussion and activities have been framed by the opportunities and concerns posed by advances in AI. These include efforts by governments of the United States, the United Kingdom, and the European Union to call for study and regulation. In October 2023, an extensive US Presidential Executive Order on *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* called for study and actions to address the possibilities of AI technologies to "exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security."[70] Directions forward for realizing the benefits of AI while minimizing risks will require continuing investments in understandings and innovation on the technical, sociotechnical, and regulatory fronts.

Conclusion

The journey of AI to date has involved decades of innovation with empirical studies and prototypes, the development of theoretical principles, and shifts among paradigms. In our overview, we shared a fast-paced arc through the history of AI as a distinct field of scientific inquiry. This journey saw a pivotal shift from early symbolic logic to probabilistic models in the mid-1980s as a response to the complexity of real-world problems. The growth and impact of the field over the last 20 years has been based largely on advancements in machine learning with efforts in discriminative models, which excel in pattern recognition and classification, and generative models, which replicate and innovate with data generation processes. The recent inflections in progress have come with advances in deep learning, which have become the foundation of today's AI applications. The current landscape of AI is defined by two significant inflection points: the rise of deep learning, and now the advent of generative AI, demonstrating both specialist and generalist competencies.

With all the rising capabilities—sprinkled with both systematic and poorly understood weaknesses—that we now see, we have little understanding of large generative AI models. There are tremendous opportunities ahead for advancing the science of AI. At the same time, we see unprecedented possibilities ahead via AI advances for leveraging computing technologies in a multitude of areas, including key domains of the biosciences, health care, the physical sciences, education, and climate and sustainability.

Notes

1. Alan M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction," *Proceedings of the London Mathematical Society* s2-43, no. 1 (1938): 544–546, https://doi.org/10.1112/plms/s2-43.6.544.

2. John Von Neumann, *First Draft of a Report on the EDVAC* (Moore School of Electrical Engineering, University of Pennsylvania, 1945), https://doi.org/10.5479/sil.538961 .39088011475779; Alan M. Turing, "Proposals for Development in the Mathematics Division of an Automatic Computing Engine (ACE)," National Physical Laboratory (NPL) (1945),

https://www.npl.co.uk/getattachment/about-us/History/Famous-faces/Alan-Turing/turing-proposal-Alan-LR.pdf?lang=en-GB.

3. Warren S. McCulloch and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5, no. 4 (December 1943): 115–133, https://doi.org/10.1007/bf02478259.

4. Norbert Wiener, *Cybernetics: Control and Communication in the Animal and the Machine*, 1st ed. (New York: Wiley, 1948).

5. John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine* 27, no. 4 (December 15, 2006): 12–12, https://doi.org/10.1609/aimag.v27i4.1904.

6. George A Miller, "The Cognitive Revolution: A Historical Perspective," *Trends in Cognitive Sciences* 7, no. 3 (March 2003): 141–144, https://doi.org/10.1016/s1364-6613(03)00029-9; Maarten Sap et al., "Quantifying the Narrative Flow of Imagined versus Autobiographical Stories," *Proceedings of the National Academy of Sciences* 119, no. 45 (November 8, 2022): e2211715119–e2211715119, https://doi.org/10.1073/pnas.2211715119.

7. See, for example, Frank Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, 1962).

8. See, for example, Allen Newell, J. C. Shaw, and Herbert A. Simon, "Report on a General Problem-Solving Program," *Semantic Scholar* (1959): 256–264, https://www.semanticscholar.org/paper/Report-on-a-general-problem-solving-program-Newell-Shaw/97876c2195ad9c7a4be010d5cb4ba6af3547421c.

9. See, for example, Bruce G. Buchanan and Edward H. Shortliffe, eds., *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*, 1st ed. (Reading, Mass: Addison-Wesley, 1984).

10. Eric J. Horvitz, John S. Breese, and Max Henrion, "Decision Theory in Expert Systems and Artificial Intelligence," *International Journal of Approximate Reasoning* 2, no. 3 (July 1988): 247–302, https://doi.org/10.1016/0888-613x(88)90120-x.

11. Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 1st ed. (San Francisco, Calif: Morgan Kaufmann, 1988).

12. See, for example, Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA: MIT Press, 2009), https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/.

13. D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward Normative Expert Systems: Part I, The Pathfinder Project," *Methods of Information in Medicine* 31, no. 02 (1992): 90–105, https://doi.org/10.1055/s-0038-1634867.

14. See, for example, Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 1998).

15. See, for example, Paul Smolensky and Geraldine Legendre, *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, vol. 1, *Cognitive Architecture* (Cambridge, MA: MIT Press, 2006); Paul Smolensky and Geraldine Legendre, *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, vol. 2, *Linguistic and Philosophical Implications* (Cambridge, MA: MIT Press, 2006); Jiayuan Mao et al., "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision," *Open Review* (2018), https://openreview.net/forum?id=rJgMlhRctm.

Systems, vol. 25 (Curran Associates, Inc., 2012), 1097–1105, https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

28. Andre Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* 542, no. 7639 (February 2, 2017): 115–118, https://doi.org/10.1038/nature21056.

29. Hinton et al., "Deep Neural Networks for Acoustic Modeling," 82–97.

30. Alex Wang et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Association for Computational Linguistics, 2018), https://doi.org/10.18653/v1/w18-5446.

31. Nestor Maslej et al., "Artificial Intelligence Index Report 2024," *arXiv*, May 29, 2024, https://doi.org/10.48550/arXiv.2405.19522.

32. David Silver et al., "Mastering the Game of Go Without Human Knowledge," *Nature* 550, no. 7676 (October 2017): 354–359, https://doi.org/10.1038/nature24270.

33. Jenna Wiens, John Guttag, and Eric Horvitz, "A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital-Specific Predictions," *JAMIA* 21, no. 4 (2014): 699–706, https://doi.org/10.1136/amiajnl-2013-002162.

34. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," *arXiv*, July 12, 2022, https://doi.org/10.48550/arXiv.2108.07258.

35. Liunian Harold Li et al., "Symbolic Chain-of-Thought Distillation: Small Models Can Also 'Think' Step-by-Step," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, vol. 1, *Long Papers* (Association for Computational Linguistics, 2023), https://doi.org/10.18653/v1/2023.acl-long.150.

36. Karan Singhal et al., "Large Language Models Encode Clinical Knowledge," *Nature* 620, no. 7972 (August 2023): 172–180, https://doi.org/10.1038/s41586-023-06291-2.

37. Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017), https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

38. Devlin et al., "BERT," 4171–4186.

39. Jared Kaplan et al., "Scaling Laws for Neural Language Models," *arXiv*, January 22, 2020, https://doi.org/10.48550/arXiv.2001.08361; Jordan Hoffmann et al., "Training Compute-Optimal Large Language Models," *arXiv*, March 29, 2022, https://doi.org/10.48550/arXiv.2203.15556.

40. Jason Wei et al., "Emergent Abilities of Large Language Models," *arXiv*, October 26, 2022, https://doi.org/10.48550/arXiv.2206.07682.

41. Michal Kosinski, "Evaluating Large Language Models in Theory of Mind Tasks," *arXiv*, February 16, 2024, https://doi.org/10.48550/arXiv.2302.02083.

42. Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv*, April 13, 2023, https://doi.org/10.48550/arXiv.2303.12712.

43. Bubeck et al., "Sparks of Artificial General Intelligence"; Karthik Valmeekam et al., "On the Planning Abilities of Large Language Models—A Critical Investigation," in *Advances in Neural Information Processing Systems*, vol. 36 (2023), 75993–75995, https://papers.nips.cc/paper_files/paper/2023/hash/ef b2072a358cef b75886a315a6fcf880 -Abstract-Conference.html.

44. Li et al., "Symbolic Chain-of-Thought Distillation: Small Models Can Also 'Think' Step-by-Step"; Mert Yuksekgonul et al., "Attention Satisfies: A Constraint-Satisfaction Lens

on Factual Errors of Language Models," *arXiv*, April 17, 2024, https://doi.org/10.48550/arXiv .2309.15098.

45. Souradip Chakraborty et al., "BioMedBERT: A Pre-Trained Biomedical Language Model for QA and IR," *Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics*, 2020, https://doi.org/10 .18653/v1/2020.coling-main.59.

46. Robert Tinn et al., "Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing," *Patterns* 4, no. 4 (April 14, 2023): 100729–100729, https://doi .org/10.1016/j.patter.2023.100729.

47. Nori et al., "Can Generalist Foundation Models Outcompete Special-Purpose Tuning?"

48. Mostafa Abdou et al., "Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color," *Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics* (2021), https://doi .org/10.18653/v1/2021.conll-1.9.

49. Chris Olah et al., "Zoom In: An Introduction to Circuits," *Distill* 5, no. 3 (March 10, 2020), https://doi.org/10.23915/distill.00024.001; Yi Zhang et al., "Unveiling Transformers with LEGO: A Synthetic Reasoning Task," *arXiv*, February 17, 2023, https://doi.org/10.48550 /arXiv.2206.04301; Catherine Olsson et al., "In-Context Learning and Induction Heads," *arXiv*, September 23, 2022, https://doi.org/10.48550/arXiv.2209.11895.

50. Yuksekgonul et al., "Attention Satisfies."

51. Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum, "Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines," *Science* 349, no. 6245 (July 17, 2015): 273–278, https://doi.org/10.1126/science.aac6076.

52. Valmeekam et al., "On the Planning Abilities of Large Language Models."

53. Eric Horvitz, "Principles and Applications of Continual Computation," *Artificial Intelligence* 126, no. 1–2 (February 2001): 159–196, https://doi.org/10.1016/s0004-3702(00)00082-5.

54. Nicholas Roy et al., "From Machine Learning to Robotics: Challenges and Opportunities for Embodied Intelligence," *arXiv*, October 28, 2021, https://doi.org/10.48550/arXiv .2110.15245.

55. T. Mitchell et al., "Never-Ending Learning," *Communications of the ACM* 61, no. 5 (April 24, 2018): 103–115, https://doi.org/10.1145/3191513.

56. Suriya Gunasekar et al., "Textbooks Are All You Need," *arXiv*, October 2, 2023, https://doi.org/10.48550/arXiv.2306.11644.

57. Sébastien Bubeck and Mark Sellke, "A Universal Law of Robustness via Isoperimetry," *Journal of the ACM* 70, no. 2 (March 21, 2023): 1–18, https://doi.org/10.1145/3578580.

58. Gunasekar et al., "Textbooks Are All You Need"; Arindam Mitra et al., "Orca 2: Teaching Small Language Models How to Reason," *arXiv*, November 21, 2023, https://doi.org /10.48550/arXiv.2311.11045.

59. Saleema Amershi et al., "Guidelines for Human-AI Interaction," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (ACM, 2019), https://doi.org /10.1145/3290605.3300233; Eric Horvitz, "Principles of Mixed-Initiative User Interfaces," *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit—CHI '99* (ACM, 1999), https://doi.org/10.1145/302979.303030.

60. Abigail Sellen and Eric Horvitz, "The Rise of the AI Co-Pilot: Lessons for Design from Aviation and Beyond," *Communications of the ACM* 67, no. 7 (June 13, 2024): 18–23, https://doi.org/10.1145/3637865.

61. Mitra et al., "Orca 2."

62. Gunasekar et al., "Textbooks Are All You Need."

63. Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," paper presented at the Conference on Language Modeling, Philadelphia, PA, October 7–9, 2024.

64. John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596 (2021): 583–589, https://www.nature.com/articles/s41586-021-03819-2.

65. Minkyung Baek et al., "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network," *Science* 373, no. 6557 (August 19, 2021): 871–876, https://www.science.org/doi/10.1126/science.abj8754.

66. Ian R Humphreys et al., "Computed Structures of Core Eukaryotic Protein Complexes," *Science* 374, no. 6573 (December 10, 2021): eabm4805–eabm4805, https://doi.org/10.1126/science.abm4805.

67. Joseph L Watson et al., "De Novo Design of Protein Structure and Function with RF-diffusion," *Nature* 620, no. 7976 (August 2023): 1089–1100, https://doi.org/10.1038/s41586-023-06415-8; Wu et al., "AutoGen."

68. Microsoft Research AI4Science and Microsoft Azure Quantum, *The Impact of Large Language Models on Scientific Discovery: A Preliminary Study Using GPT-4* (November 2023), https://arxiv.org/pdf/2311.07361.

69. Eric Horvitz et al., "Now, Later, and Lasting: Ten Priorities for AI Research, Policy, and Practice," *Communications of the ACM* 67, no. 6 (June 1, 2024): 39–40, https://dl.acm.org/doi/pdf/10.1145/3637866.

70. Joseph R. Biden, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," *Federal Register* 88, no. 210 (October 30, 2023): 75191–75226.