

# The Modern Artificial Intelligence Primer

With A Special Focus On Generative AI

# Contents

- Foreword ..... 1
- 1 What is Artificial Intelligence and Where Does Generative AI Fit? ..... 2**
  - A Timeline Toward the Rise of GenAI and Beyond .....4
- 2 Understanding A Simple Neural Network ..... 9**
  - Introducing the Multilayer Perceptron .....10
  - Entropy: Designing a Neural Network to Learn .....13
  - “Backprop”: Training a Neural Network .....13
- 3 Advances In Modern AI ..... 14**
  - 2014: Generative Adversarial Networks .....15
  - 2016: Deep Reinforcement Learning and AlphaGo .....16
  - 2020: Diffusion Models .....16
  - 2022: ChatGPT.....18
  - 2022: Multimodal AI .....20
- 4 A Modern Generative Large Language Model System .....21**
  - Large Language Models.....22
  - Fine-Tuning.....23
  - Agents.....23
  - Reinforcement Learning with Human Feedback .....24
  - Prompt Engineering .....25
  - Retrieval-Augmented Generation .....27
  - How Vector Search Works.....27
  - Accelerated Computing .....28
  - When to Use GenAI .....29
  - Creativity Versus Certainty .....31
  - AI Operations.....31
- 5 Recognizing AI’s Limits and Challenges .....32**
  - AI Bias .....33
  - Generative Hallucinations.....35
  - AI Vulnerability.....36
- 6 A Look Ahead At Modern AI’s Potential Realized .....37**
  - The Path to Artificial General Intelligence (AGI) .....39
- Acknowledgments ..... 40**
- Glossary of Terms ..... 40**
- Endnotes .....42**



“I often tell my students not to be misled by the name ‘artificial intelligence’—there is nothing artificial about it. AI is made by humans, intended to behave by humans, and ultimately, to impact humans’ lives and human society.”<sup>1</sup>

**Fei-Fei Li**

Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence

# Foreword

2024-2025

In 2017, we published our first **Artificial Intelligence Primer**. It focused on recent advances in supervised and unsupervised machine learning (ML), with a heavy emphasis on big data. At the time, many of these techniques were still being honed inside the laboratory, with academic and other research institutions leading development. Fast-forward 7 years, and we now live in a world where artificial intelligence (AI) has firmly entered the mainstream, with generative AI (GenAI), multimodal AI (MMAI), and autonomous systems poised to further disrupt government, industry, and society.

Today, the challenge for all of us is to look beyond the hype to discern the real value and science of AI systems. To aid in these assessments, we have updated our previous report to produce this Modern Artificial Intelligence Primer. Our goal is to demystify this technology and empower leaders to make smart, thoughtful decisions about AI’s development and applications. A key focus is ensuring its responsible use.

Throughout this primer, we strive to address today’s most important questions about AI, including:

- What should enterprises consider when evaluating AI?
- What technical advances finally led to the GenAI breakthrough?
- Why, exactly, are neural networks designed the way they are?
- What risks and challenges does AI pose?
- Where is AI headed next?

This primer builds upon Booz Allen’s 110 years of partnering with clients to address their most strategic challenges, 3 decades of data science leadership, and extensive record of delivering milestones in public-sector AI deployment. As a leading provider of AI services<sup>2</sup>— we are uniquely equipped to share real-world insights secured in arguably the most demanding and scrutinized proving ground for AI anywhere on the planet. Our broad and deep experience encompasses more than 200 AI services engagements with more than 160 clients, including large AI contracts within the U.S. Department of Defense (DOD). By bringing these advanced perspectives together, we hope to create a clearer picture of AI for organizations everywhere.

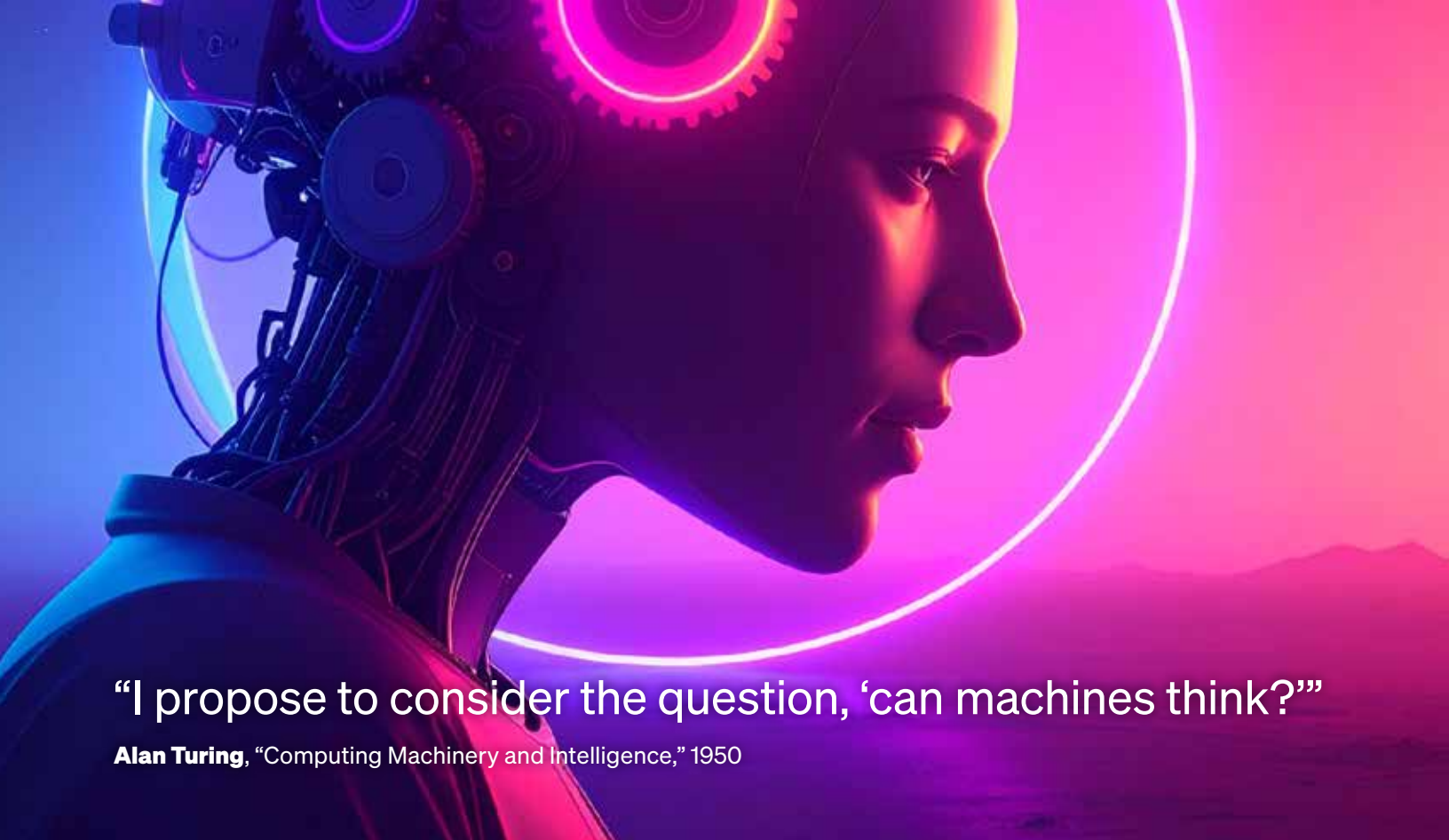
While the advances in AI over the past decade have been amazing, we also believe future innovations will be even more impactful. Harnessed responsibly, they will help us grow prosperity, improve healthcare outcomes, transition to a more sustainable future, and maintain global security. Collaborating closely with our clients and partners, we are committed to being at the forefront of shaping this better future.

**John Larson**

Executive Vice President  
Booz Allen Hamilton Chief Technology Office

# **What Is Artificial Intelligence And Where Does Generative AI Fit?**





**“I propose to consider the question, ‘can machines think?’”**

**Alan Turing**, “Computing Machinery and Intelligence,” 1950

Let’s begin with a basic question: what do we mean by “artificial intelligence”? In its simplest form, AI allows us to create systems and machines that can perform tasks and answer questions that would normally require human intelligence. AI systems receive data as an input and produce some desired output—and the different processes for making that output define the different AI techniques.

Underpinning much of AI, ML encompasses the mathematical techniques used to automate pattern recognition, with ML algorithms being given many examples of correct outputs for specific inputs to try to mimic the process. AI includes traditional ML algorithms, like logistic regression, decision trees, and support vector machines, along with neural networks in the field of deep learning, like convolutional neural networks (CNN) and transformers.

Researchers developed these methods to encode high-level strategies for how to solve problems. Common AI capabilities support applications such as virtual assistants, facial recognition tools, image labeling, search and summarization, and many more. We will delve into these topics shortly.

This leads us to generative artificial intelligence (GenAI) specifically: GenAI is a type of AI focused on systems that can interpret commands and dynamically generate

appropriate responses in the form of text, images, audio and video, software code, music scores, and digital designs and models, among other outputs. It encompasses not only large language models (LLM) that power virtual assistants such as OpenAI’s ChatGPT, but also image algorithms that power applications like Midjourney, and increasingly, video generators like OpenAI’s Sora.

However, it’s important to recognize that the sum of GenAI’s impressive results would not be possible without the collection of relatively old parts that provide its foundation. For example, many of the components that inform today’s generative LLM systems are models and methods reconstituted from prior innovations—transfer learning (2012),<sup>3</sup> self-attention (2015),<sup>4</sup> transformers (2017),<sup>5</sup> and reinforcement learning algorithms like proximal policy optimization (PPO) (2017).<sup>6</sup>

This primer analyzes some of the basic mechanics of GenAI’s smaller parts to help build a greater understanding of how it works and why modern AI, in all its forms, exerts such transformative power. After all, the next big thing in AI is likely to be built from tools we are making today.

# A Timeline Toward The Rise Of Gen AI And Beyond

The year 1950 remains a pivotal one in AI's evolution. That's when mathematician and computer scientist Alan Turing introduced the Imitation Game (or "Turing Test") in his seminal paper "Computing Machinery and Intelligence."<sup>7</sup> With the idea of a conversation in which machine and human become indistinguishable to an evaluator, Turing formulated and brought urgency to a complex question that still demands consideration today: "Can machines think?" Turing predicted that by the year 2000 humans would be able to program computers to play the Imitation Game at 70% accuracy.

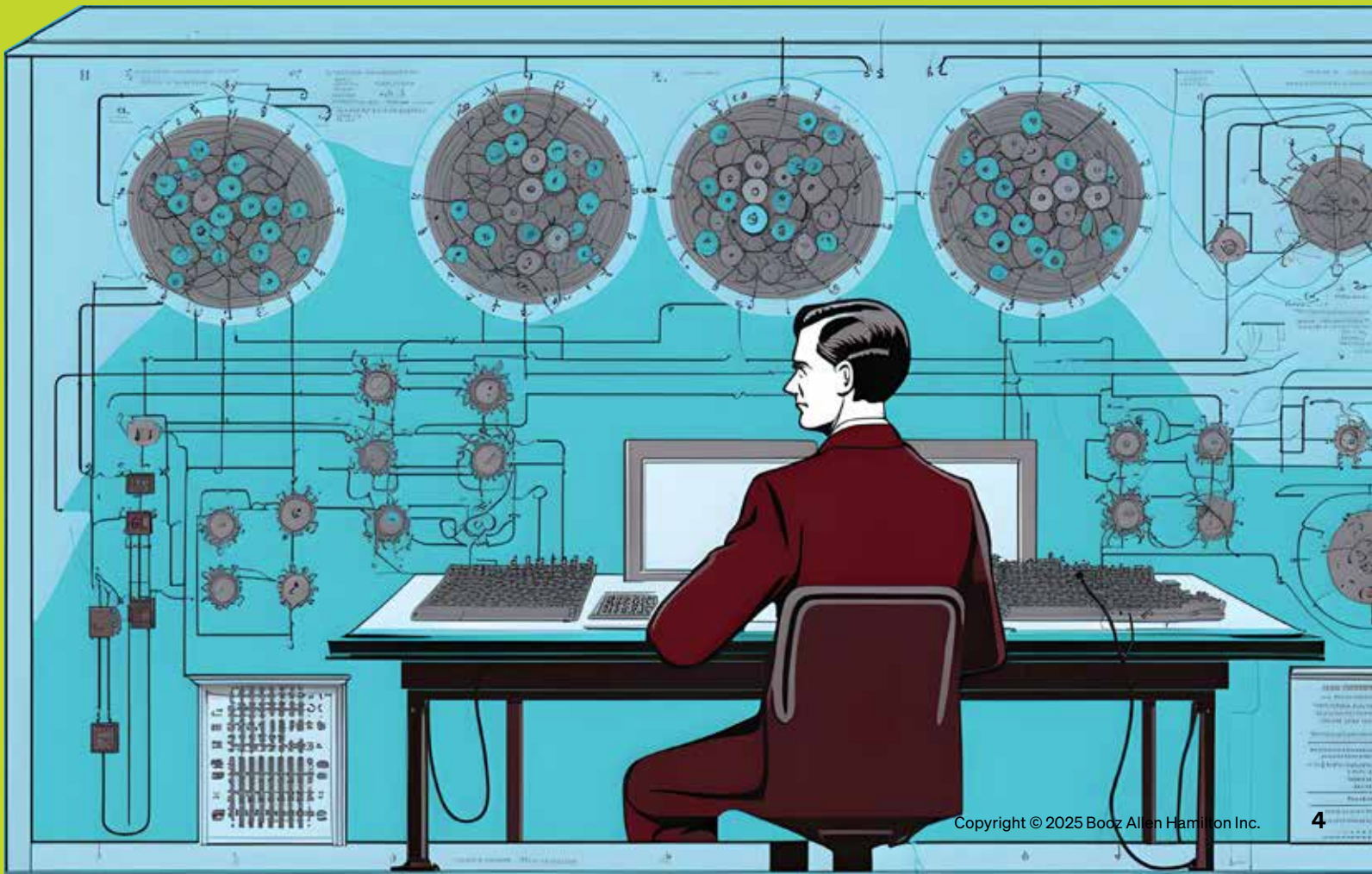
A few years later, in the summer of 1956, a small group of mathematicians and scientists met at Dartmouth

College to kick off what is widely viewed as the birth of formal AI research. Research within the field at large continued productively for more than a decade. But despite the introduction of novel methods during AI's initial Golden Age, a so-called "AI winter" set in throughout the 1970s and 1980s. During that time, AI interest, investment, and research floundered due to failures linked to machine translation, expert systems, and inefficient training algorithms.

But as the 1990s approached, breakthroughs again proliferated. A key example: In 1987, the notion of a "connectionist architecture,"<sup>8</sup> the predecessor framework to neural networks, was introduced. This was an early step in taking inspiration from biology in having multiple "layers" of processing, which increases the flexibility and capacity of a model compared to an equally large but "shallow" (i.e., one-layer) model.

## Alan Turing

"Can machines think?" Turing predicted that by the year **2000** humans would be able to program computers to play the Imitation Game at **70%** accuracy.





Neural networks are the fundamental types of algorithms used in deep learning. Researchers found that an AI model could “learn” when they updated its weights—and gradually reduced the difference between the model’s predicted and demonstrated outcomes—through the technique of backpropagation, an error-elimination method introduced a year earlier. And in 1989, computer scientist Yann LeCun introduced the first real-world implementation of what he would later term a “convolutional neural network (CNN)”<sup>9</sup> on U.S. Postal Service data for handwritten digit recognition.<sup>10</sup>

With momentum gathering, other advances soon emerged that, together, helped drive and accelerate AI’s maturation and evolution:

In 1997, IBM’s Deep Blue system clearly demonstrated AI’s potential to learn when it beat world champion Garry Kasparov at chess through the use of a supercomputer to rapidly search for and identify optimal moves. This began to spark the notion that computers can learn on their own, made even more powerful when AlphaGo beat the leading human champion in the game of Go.

In the same year, researchers revolutionized the field of natural language processing (NLP) with the introduction of the long short-term memory (LSTM)<sup>11</sup> algorithm (Figure 1). This was important because it enabled the recurrent neural

network (RNN) algorithm to more effectively “remember” and “forget.” Its use case was a predecessor to the algorithms we see now for LLMs.

In 2012, research in computer vision expanded with the creation of AlexNet, an eight-layer CNN designed by the University of Toronto’s Alex Krizhevsky and Ilya Sutskever that introduced the ReLU non-linearity. It was trained on the now-famous ImageNet dataset and won the ImageNet Large Scale Visual Recognition Challenge that year with a top-5 error rate of 15.3%, more than 10 percentage points below the next competitor. The CNN is still used as the de facto computer vision algorithm to detect objects and classify and segment images.

Another advance in NLP followed the next year, as Google published the word2vec algorithm,<sup>12</sup> a so-called “skip-gram” model that can make predictions about how to complete partial sentences through its analysis of the surrounding words and their meanings. Models like word2vec introduced the idea of an “embedding” that is now a key component of all LLMs and that compresses the value of text information in vector representations.

And in 2014, the advent of generative adversarial networks (GAN), which produce outputs that replicate the exact characteristics of a given set of training data, opened new possibilities but also risks related to AI image generation—including the rapid production of

highly convincing “deepfakes” that seek to fool audiences with manipulated content.

Where are we now? As time passed, researchers developed thousands of variations of these algorithms. These contributions include more efficient ways for the neural network to learn, such as different functions to minimize error; novel layers, which introduce stability during training or increase computational speed; and setups that capture complex dependencies and representations within the data. Most recent advances in AI rely on innovative deep learning techniques using these neural networks, which we explore in more detail in the following sections.

Ultimately, all of this presents a paradox: AI continues to build upon (and accelerate) previous advances while also reflecting new approaches. In selecting use cases and applications, enterprises will need to weigh AI’s traditional deterministic approach against GenAI’s increasingly probabilistic nature. The latter enables the greater creativity and independence that have excited so many, albeit at a potential cost of assurance and control. As we will discover, modern AI requires new thinking on how to apply and safeguard the technology to augment and enhance human performance while minimizing harmful disruption and anxiety.



# From AI's Golden Age To The GenAI Era

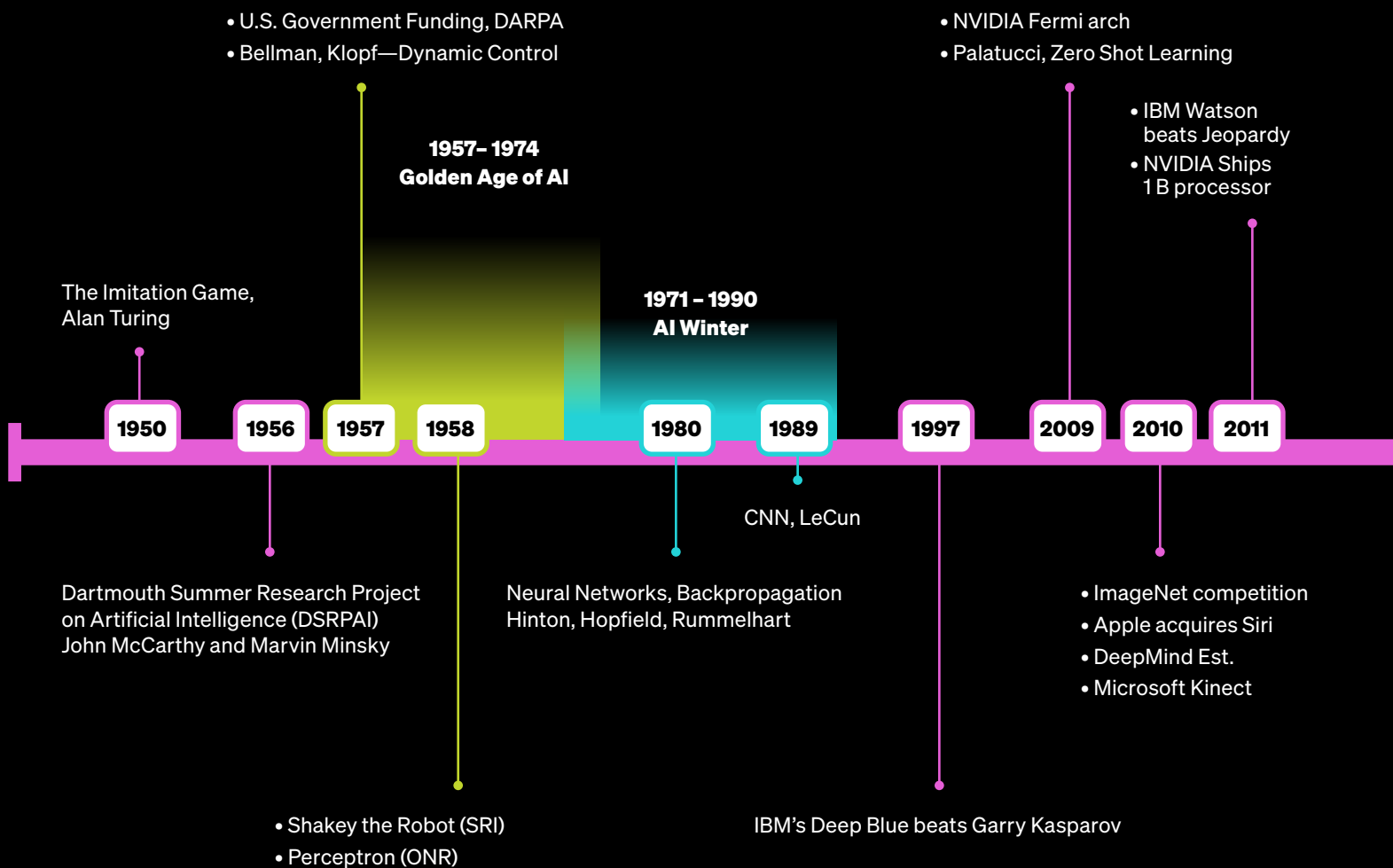
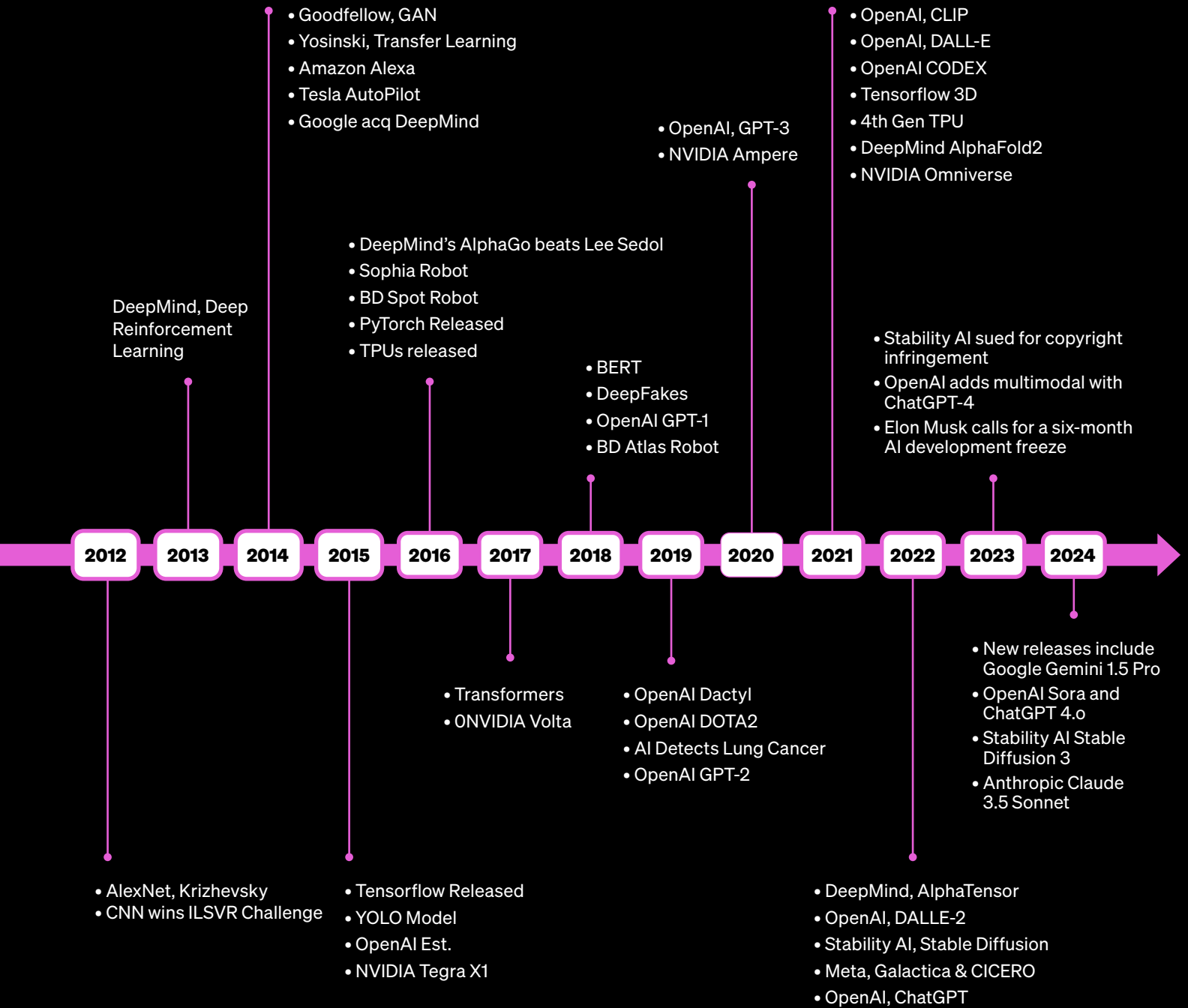
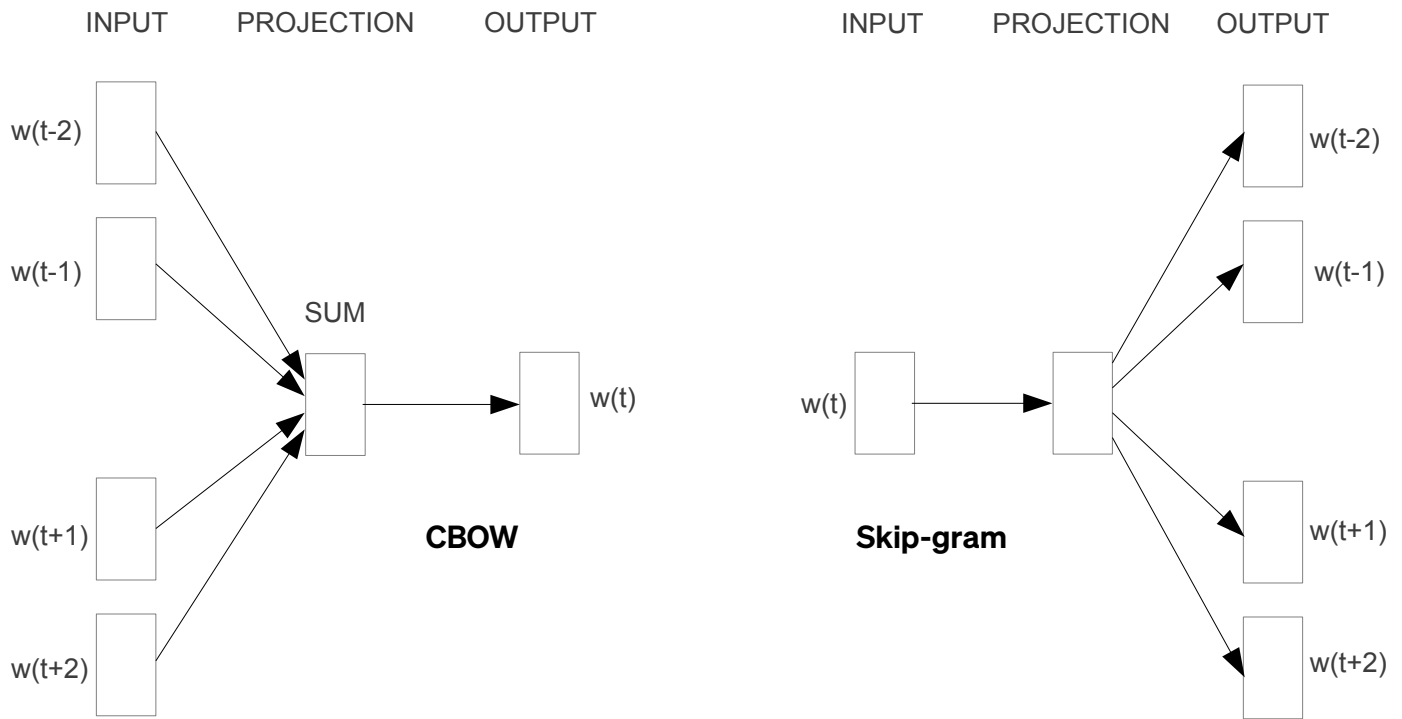


Figure 2: From AI's Golden Age To The GenAI Era





## Skip-Gram Model



## LSTM

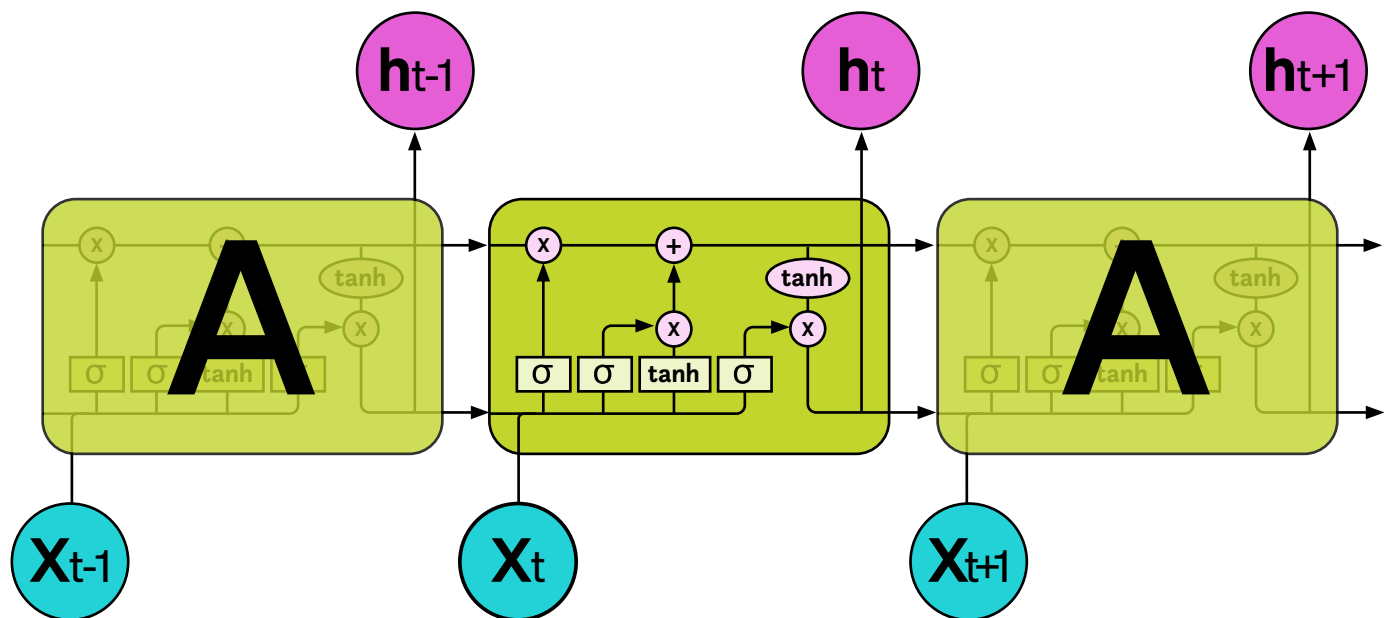


Figure 1: NLP Innovations<sup>13,14</sup>

# Understanding A Simple Neural Network

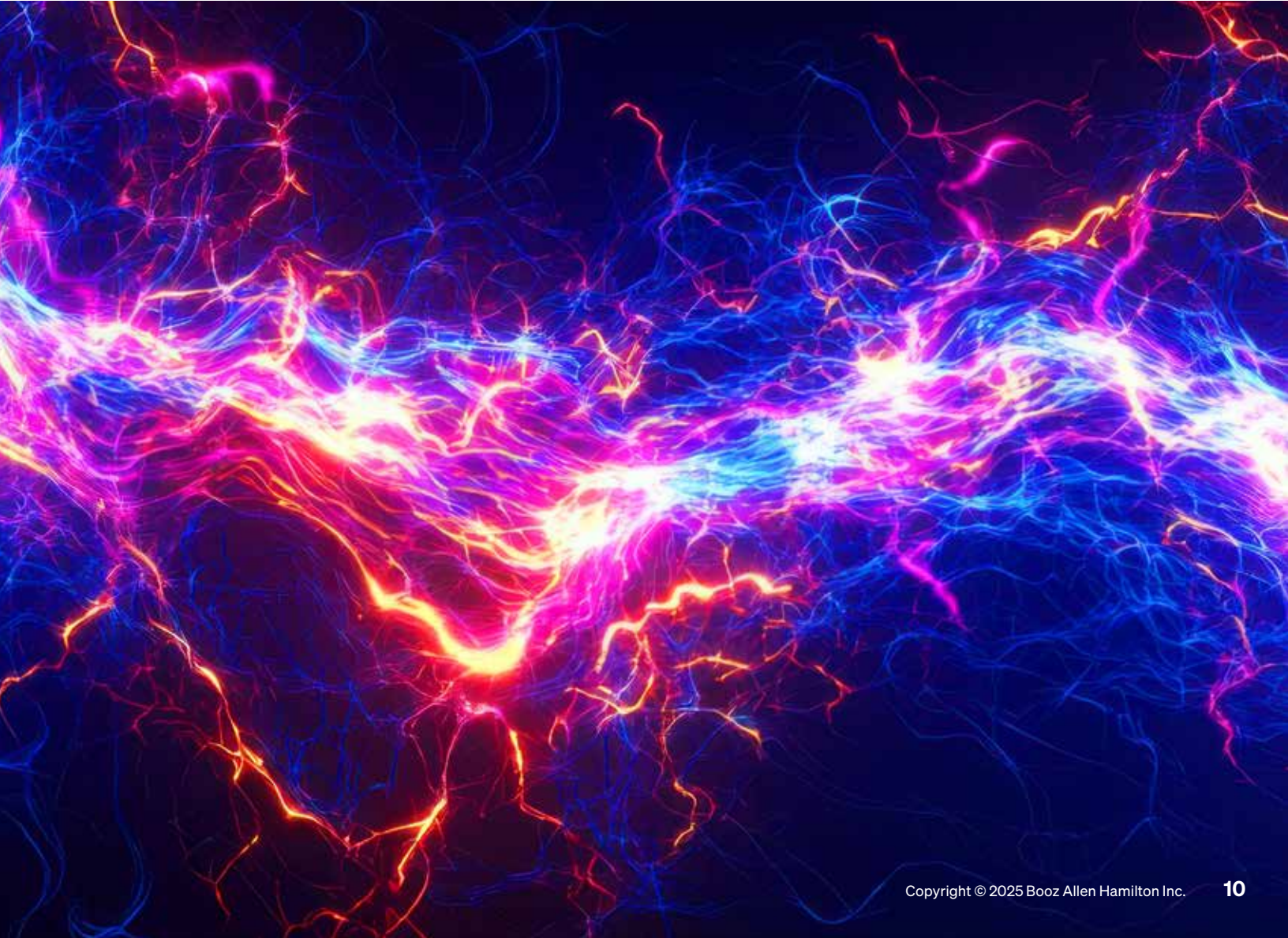
Mimicking the structure of the human brain, a neural network is a key part of the specific ML technique called “deep learning.” Neural networks allow systems to approximate non-linear functions used to produce complex calculations involving highly variable relationships and outcomes that appear highly unpredictable. Understanding non-linear relationships is important as many facets of the real world are governed by often unknown factors.

We’ll examine the technical underpinnings of neural networks using the simplest interpretation of such networks: a “multilayer perceptron” (MLP) or “feedforward neural network.” Here’s why we’re taking a closer look: Although the MLP is a fairly simple example of a deep learning algorithm, it is powerful and utilitarian and is used so much in AI that it has become a key algorithm in the backbone of LLMs like Anthropic’s Claude.

## Introducing The Multilayer Perceptron

When we architect neural networks—that is, the algorithms that enable deep learning—we have a variety of design choices to consider (Figure 3):

- **Recurrent Neural Networks** (RNN) focus on sequence-to-sequence inputs, such as sentences.
- **Convolutional Neural Networks** (CNN) apply convolution that uses multidimensional filters to learn image maps.
- **Generative Adversarial Networks** (GAN) provide frameworks where two or more neural networks compete against each other to accomplish tasks, such as generating synthetic training data.
- **Diffusion Models** rely on a traditional U-Net<sup>15</sup> neural network—which can harness limited data without losing speed and precision—for generating images and other media.





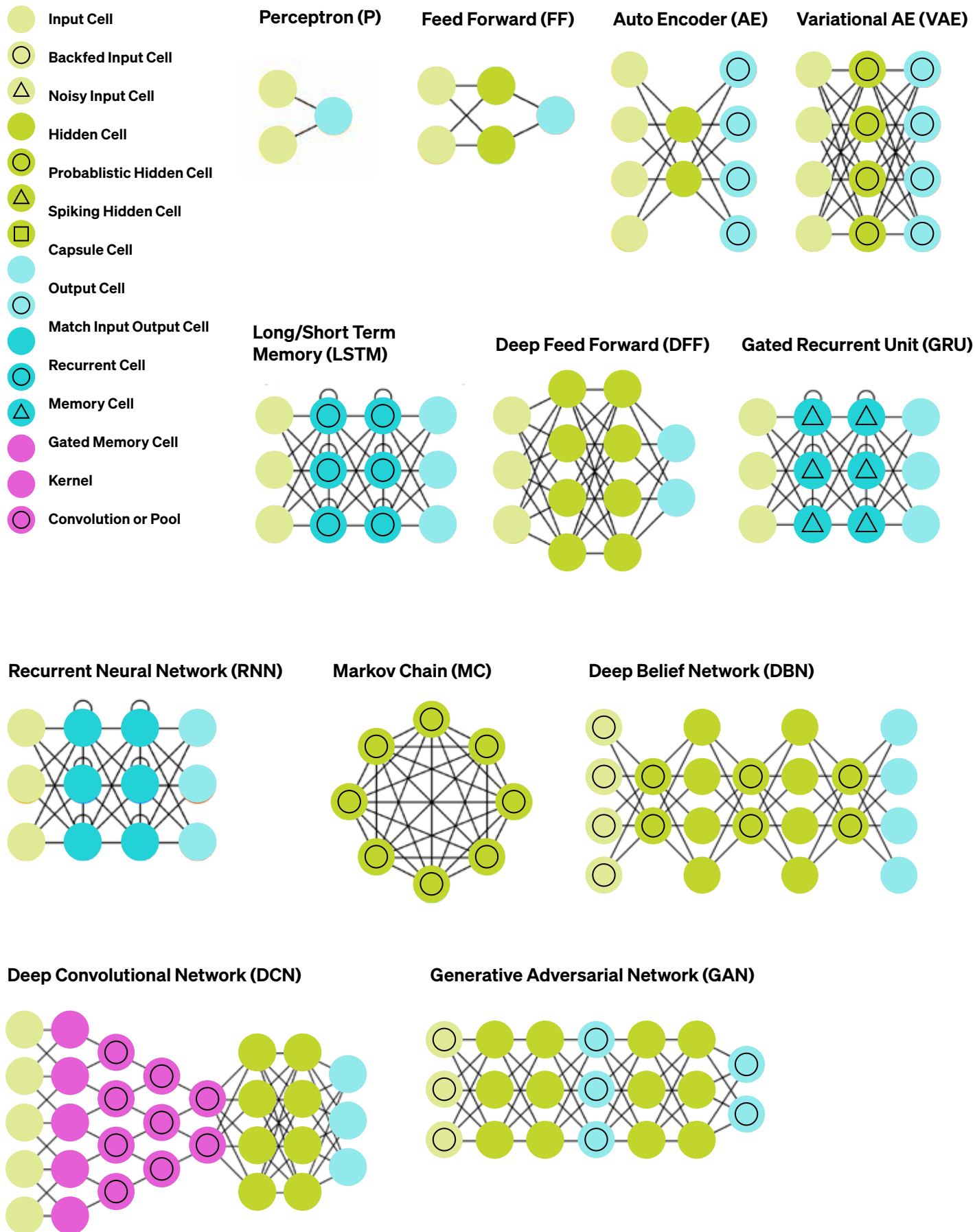


Figure 3: Model Zoo<sup>16</sup>

Regardless of the apparent complexity of the neural network we choose, it is always just a large grouping of many smaller functions—essentially, “a small set of highschool-level ideas put together.”<sup>17</sup> For example, in a basic neural network, the functions that perform low-level decision making at the feature-map level are called “non-linearities.” They are embedded within an equation to independently perform linear regression calculations. When combined in a neural network, these complex sets of linear regression equations can be stitched together to learn representations of images and sentences (Figure 4).

MLPs often perform complex data classification and pattern recognition tasks. In the MLP depicted here, multiple composite functions are used to make a

prediction  $\hat{y}$ . Notably in this composition, the weights outputted from each layer ( $W^1, W^2, W^3$ ) are the inputs to the next layer, like a loop fed into itself. Therefore,  $W^3 \sigma$  is a function of  $W^2 \sigma$  which itself is a function of  $W^1 \sigma$ , where the symbol  $\sigma$  represents the sigmoid (non-linearity) function—the well-known “S” curve—denoted as “a” for activation.

Composition functions are simply a more formal way to explain what is often called “modularity.”<sup>18</sup> Modularity is the ability to swap different non-linearities in and out, adapt the architecture for different layers and connectivity patterns, apply special blocks, and add regularization, normalization, and more.

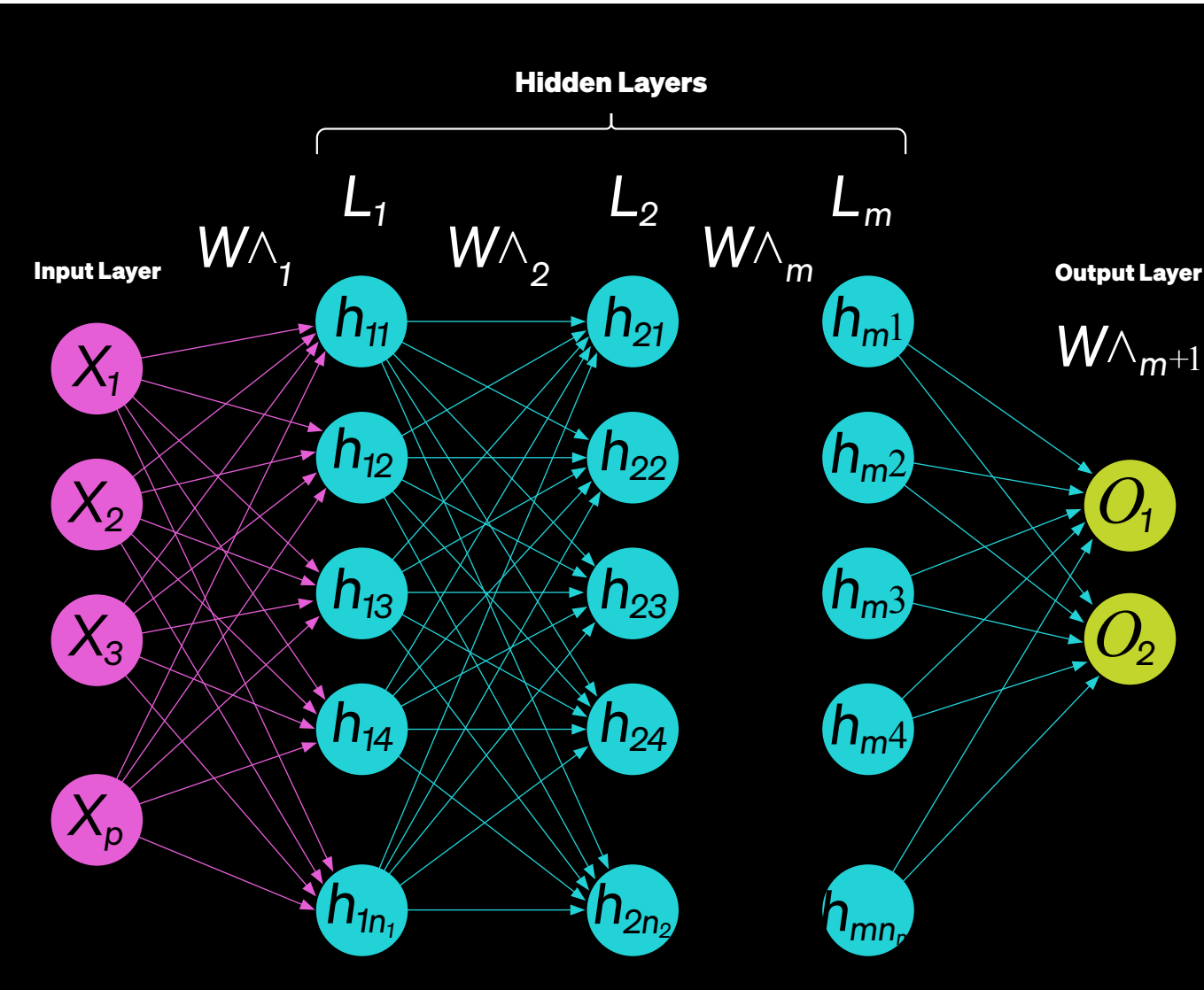


Figure 4: A Feedforward Neural Network (Courtesy Of Quebec Ai)<sup>19</sup>

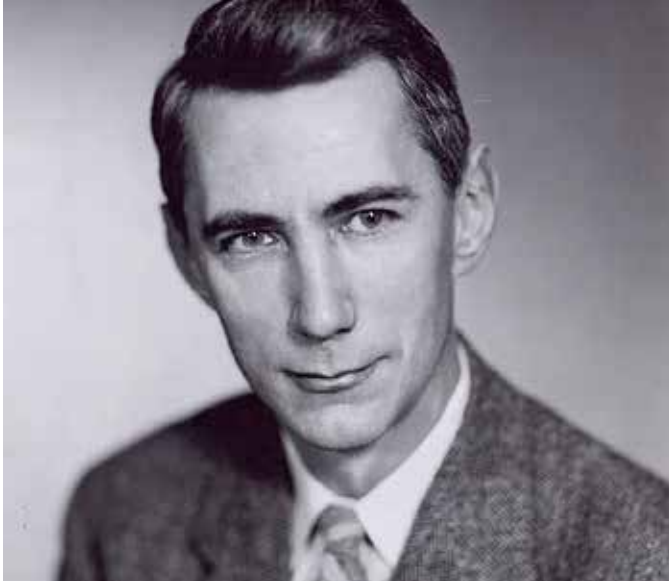


Figure 5: Claude Shannon

## Entropy: Designing A Neural Network To Learn

An important aspect of these activation functions is that they are differentiable, meaning that their derivatives—their sensitivity to changes in the input data—can be calculated to guide the neural network toward making more accurate predictions. For example, we can fine-tune the activation function to better fit the data by searching the space of all possible parameters or weights within the neural network. Gradient descent is a well-known optimization algorithm that minimizes errors by examining “loss functions”—formulas that minimize the difference between predicted and actual outcomes—to refine weighting.

One example of a well-known loss function used to train neural networks is cross-entropy. In his 1948 paper “A Mathematical Theory of Communication,”<sup>20</sup> researcher Claude Shannon (Figure 5) positioned information entropy, a precursor of cross-entropy, as a measure of uncertainty or randomness. A very “surprising” event rarely happens and therefore has a very low probability of occurring (e.g., 2%). On the other hand, an “unsurprising” event, one that happens frequently, has a high probability (e.g., 95%). For a single event, we can calculate its *entropy*. While a highly probable event like the Sun rising has low entropy (almost no surprise), an improbable event like two Suns rising would have very high entropy.

Within our simple neural network, we can measure the average entropy across all events, or the *cross-entropy*<sup>21</sup> of a model on some data distribution. A correct model needs to be the least surprised by the unseen data. It should have low entropy because the algorithm has mapped all the feature distributions and has achieved good generalization. It “knows” unfamiliar events so well that it can expect them to occur. With entropy minimized, errors are less common and learning becomes more efficient.

## “Backprop”: Training A Neural Network

We can use the errors calculated through the loss function as a source of feedback during backpropagation (sometimes shortened to “backprop”)—an error-assignment process that has been usefully compared to “a Kafkaesque judicial system” characterized by “rounds of recursive fingerprinting.”<sup>23</sup>

In backpropagation,<sup>24</sup> the algorithm runs backward from output to input, assigning “blame” for the final error to preceding steps. This involves the use of calculus to compute the weights’ respective errors, layer by layer, where changes to the algorithm are calculated as gradients. Engineers can inspect the gradients to uncover an essential truth—whether the algorithm is learning or not. For example, if the gradients are consistently close to zero, we can say they are “saturated” and that the algorithm isn’t learning (Figure 6).

A simple neural network like this MLP relies on one loss function to determine how well the model fits the data. Changing the loss function can allow the network to learn different kinds of problems. But even the simplest neural networks provide the underlying foundation for many of the most transformational achievements of 21st-century AI.

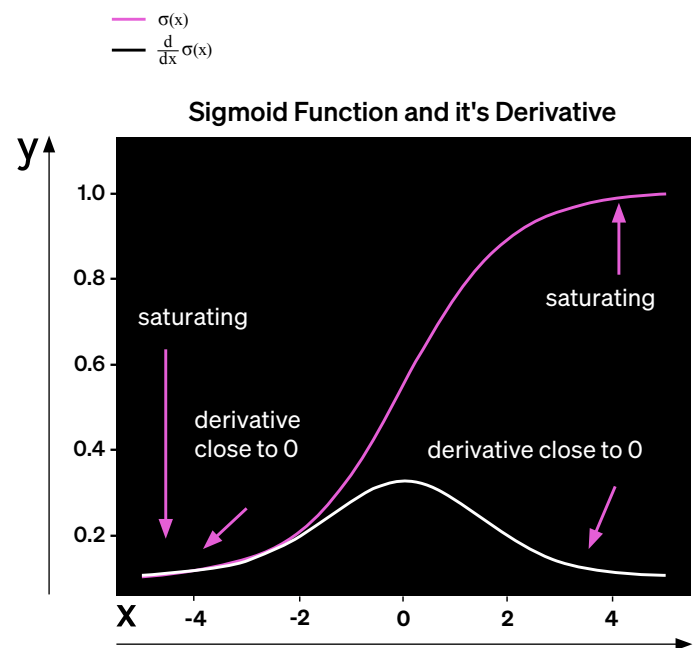


Figure 6: Gradients of a Sigmoid Function<sup>22</sup>

# Advances In Modern AI







Over the past decade, progress in AI research has accelerated dramatically, with significant advances shaping the modern AI that dominates today's headlines. We profile several of the most influential here: GANs, deep reinforcement learning (DRL), diffusion models, ChatGPT, and MMAI. Many of these advances rely on a new way to encode the idea of “self-supervision:” that the loss function (what the model learns) can be specified using only the original data—with no human annotations or input required. Combining self-supervision on an enormous amount of data with human supervision on a small amount of data has been the key strategy for many of these advancements.

## 2014: Generative Adversarial Networks

Introduced by Ian Goodfellow in 2014, a GAN can create new data instances resembling the training data used to

develop it, such as text or images. GANs combine two functions—a generator and a discriminator—as neural networks (typically, or as any differentiable function) in a cat-and-mouse game that continually refines performance.

Specifically, the generator strives to create new, prompt-inspired data that is indistinguishable in quality from the training data. Meanwhile, the discriminator works to identify this so-called fake or generated data. The back-and-forth competition between the two functions is an adversarial process in which the output becomes increasingly realistic and independent of actual data (Figure 7).

Through this process, GANs can learn high-dimensional mappings of complex data like videos, images, and text. For this reason, GANs were an important step toward the fully realized image- and text-generation capabilities of GenAI.

**“G” is the generator and “D” is the discriminator.  
Both are neural networks.**

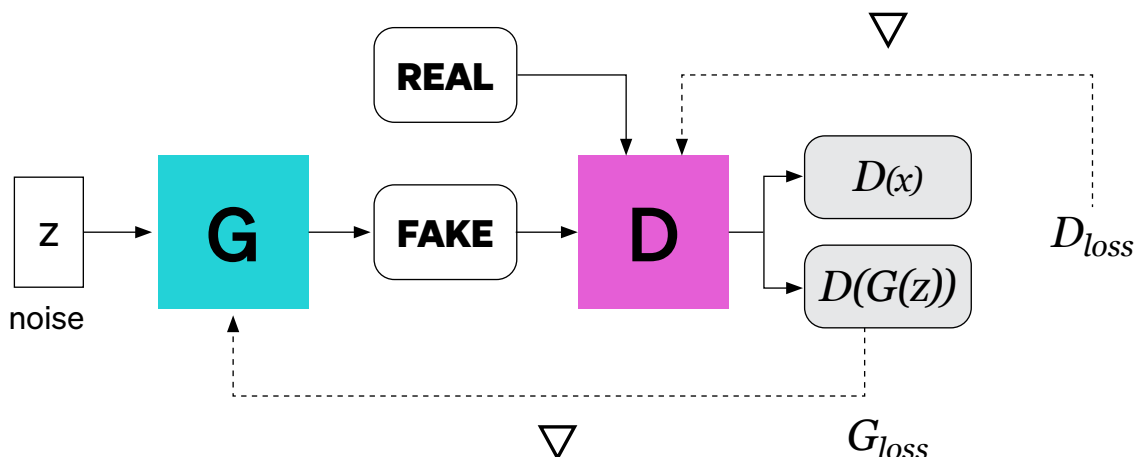


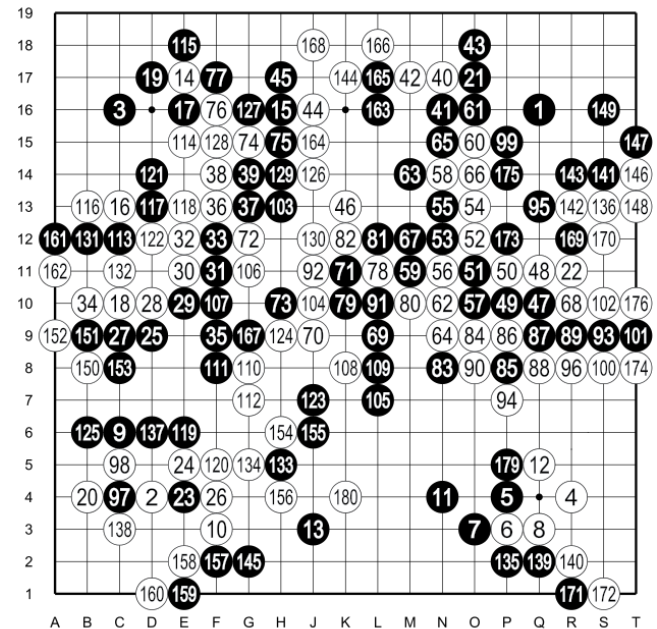
Figure 7: The Gan Algorithm<sup>25</sup>

## 2016: Deep Reinforcement Learning And AlphaGo

Developed by DeepMind in 2016,<sup>26</sup> DRL reflects within a single neural network the concept of learning by trial and error along with the ability to learn from direct inputs in human-like ways. The neural network of the Google DeepMind AI named “AlphaGo,” which plays the strategy board game Go, was trained to learn while playing through DRL.

On March 10, 2016, AlphaGo defeated one of the world’s top Go champions, Lee Sedol. This critical turn in the game’s 2,500-year history came when AlphaGo devised the now-famous “Move 37,” which had a 1-in-10,000 likelihood of being played by a human (Figure 8).<sup>27</sup>

AlphaGo was trained in a supervised manner from 160,000 human expert games along with DRL through games against itself, where it self-played 30 million distinct positions, each sampled from a separate game. Move 37 marked a groundbreaking moment because the AI had apparently designed its own creative strategy on the fly due to the way it was trained, crystallizing AI’s potential for ingenuity.



Lee Sedol (W) vs AlphaGo (B) - Game 4  
177 at 51 178 at 57

Figure 8: Move<sup>37</sup> [Wikimedia Commons](#)

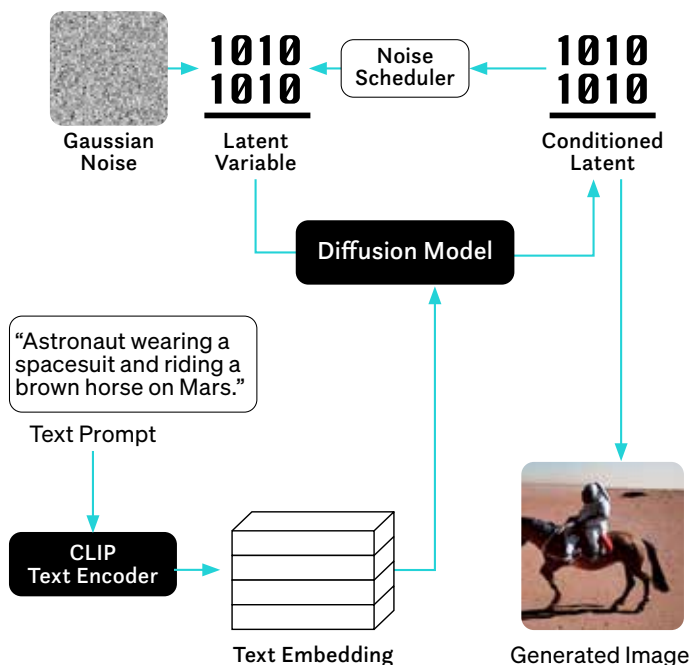


Figure 9: How AI Art Models Work

## 2020: Diffusion Models

AI art generators like Stable Diffusion, DALL-E, and Midjourney have captured the world’s imagination by turning text prompts infused with descriptions and design information into photorealistic, high-resolution imagery.

These art generators pair a language processing algorithm with an image generation algorithm to create this imagery. The language processing part uses OpenAI’s Contrastive Language-Image Pre-Training (CLIP), which has already been trained on 400 million image or text pairs. It learns to understand visual concepts from natural language, translating prompts into text tokens from the CLIP model that are passed to the diffusion model as an input supported by an attention layer within a GAN.

Diffusion models are newer approaches that can deliver more realistic results than the GANs that preceded them. This is because diffusion models are “conditioned” (Figure 9) to generate more sophisticated images, which involves adding and removing noise from the image in a controlled manner to teach the model.

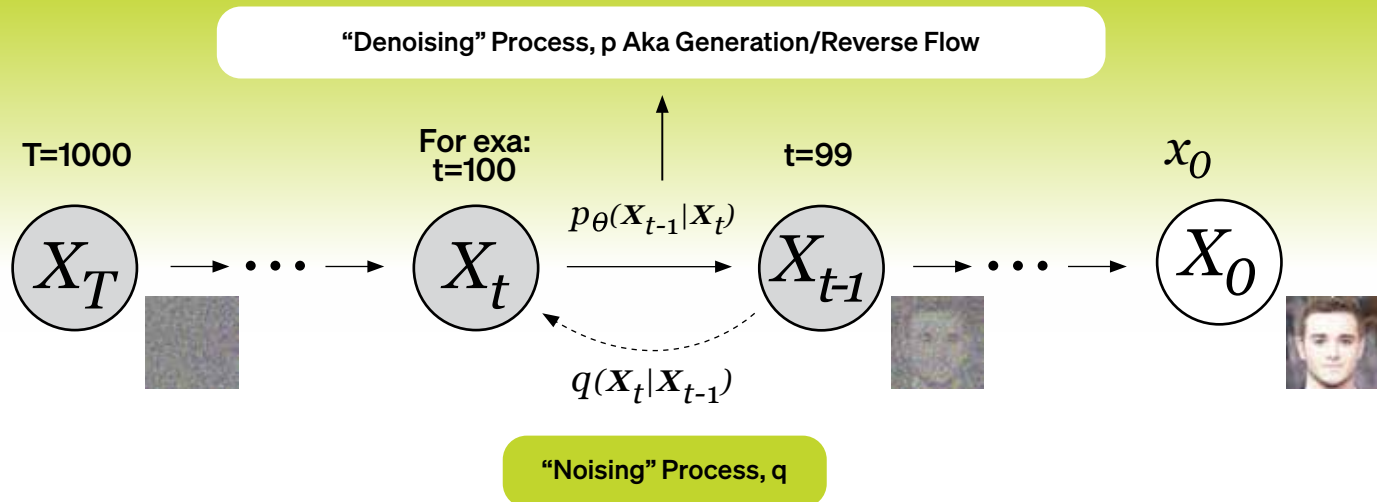


Figure 10: Diffusion Model Training Process<sup>32</sup>

Let’s look more closely at how we train diffusion models. Figure 10 shows the basic schema of a diffusion model with two core processes. In orange, there is the forward process where “Gaussian” noise is added to a “clean” image (image  $X_0$  in figure 10). This noise is gradually added to each subsequent image to destroy it over time, with a specific schedule dictating how much noise is added at each time step.

The green shows the reverse or “denoising” process that serves as the generative phase. Specifically, the diffusion model creates new imagery by removing extraneous noise, leaving only the prompted picture in its place. The AI element occurs here in the generative phase since a neural network (U-Net) is used to predict how much noise to remove at each time step, going backward.

Training diffusion models is non-trivial primarily due to the long iterative denoising process. For example, it’s estimated that training a state-of-the-art diffusion model can take 150 to 1,000 V100 days, meaning that it would take 5 days to create 50,000 samples (V100 days are a computing measure based on the output of a single NVIDIA A100 graphics processing unit [GPU]).<sup>28</sup>

Increasing the speed of generation is an active area of research focused on faster sampling strategies such as grid search and network compression,<sup>29</sup> as well as post-training quantization and pruning.<sup>30</sup> To put this into perspective, a recent study found that the cost of training frontier models has grown 2–3 times annually during the past 8 years, with the projected cost for building next-generation models reaching more than \$1 billion by 2027.<sup>31</sup>



# 2022: ChatGPT

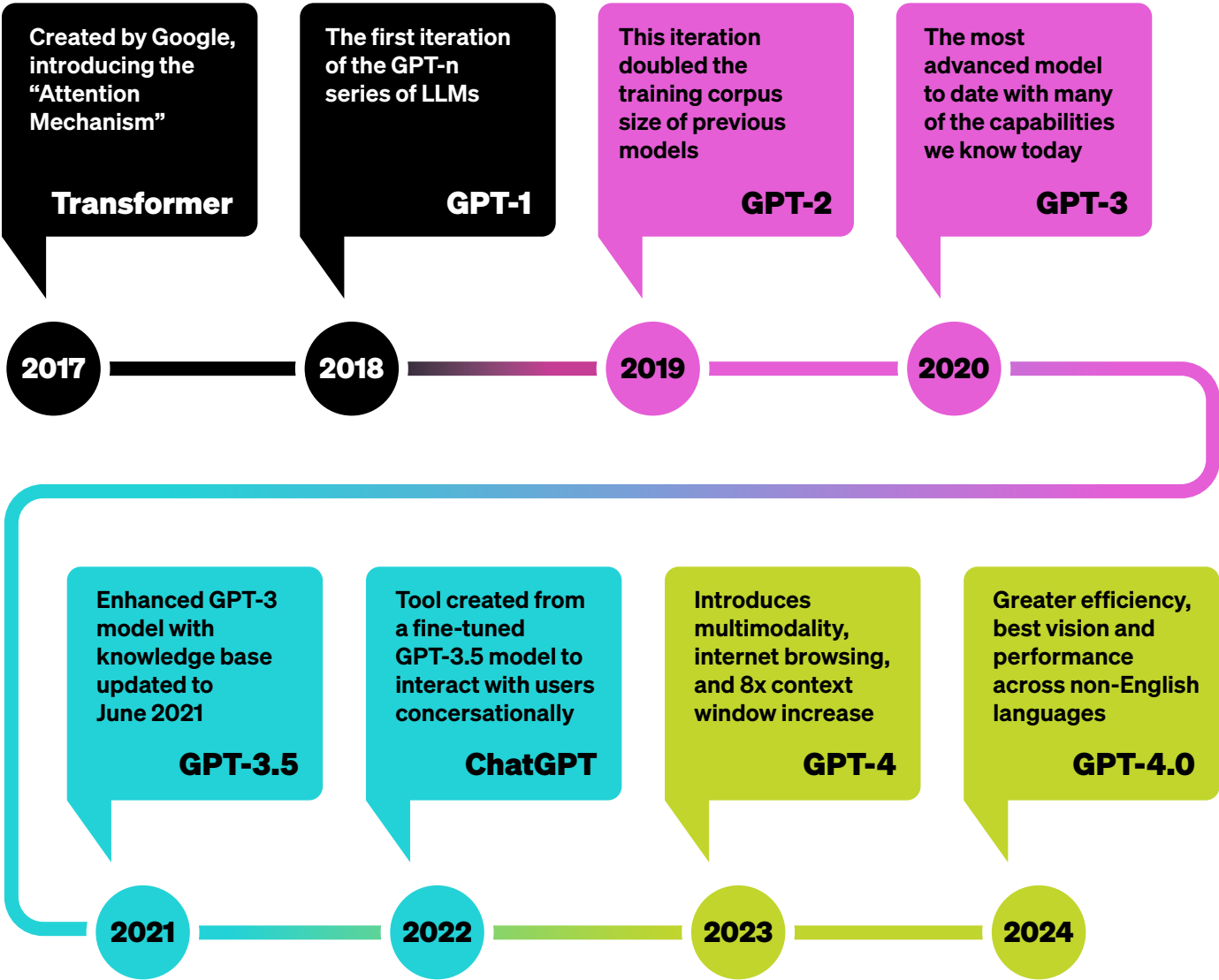
ChatGPT is a well-known AI chatbot and virtual assistant developed by San Francisco-based startup OpenAI. When it first came out, the chatbot used the GPT-3 model architecture—specifically, GPT-3.5-turbo—to generate content in response to prompts it had not been explicitly trained on. ChatGPT is the culmination of advancements from the original GPT to the current model, GPT-4.o (Figure 11).

The key advancements leading to ChatGPT involved combining older ideas into a new system based on the 2022 InstructGPT paper.<sup>33</sup> The InstructGPT models were trained to promote alignment, a concept introduced by DeepMind in 2018,<sup>34</sup> so that language

models would output responses that act in accordance with the user’s intention.

To do so, the GPT-3 base model was fine-tuned using reinforcement learning from human feedback (RLHF) developed in 2017 by OpenAI.<sup>35</sup> With RLHF, humans provide feedback on an AI system’s behavior that is used to define an AI task as opposed to crafting a manual reward function. For ChatGPT, human preference is the reward signal collected from GPT-3 outputs that are scored by humans from best to worst.

The three-step process consists of (1) a pre-trained GPT-3 model on human-demonstrated desired behavior of input prompts; (2) a supervised fine-tuning (SFT) model that trains a reward model (RM) based on human-preferred language outputs, such as instruction-output



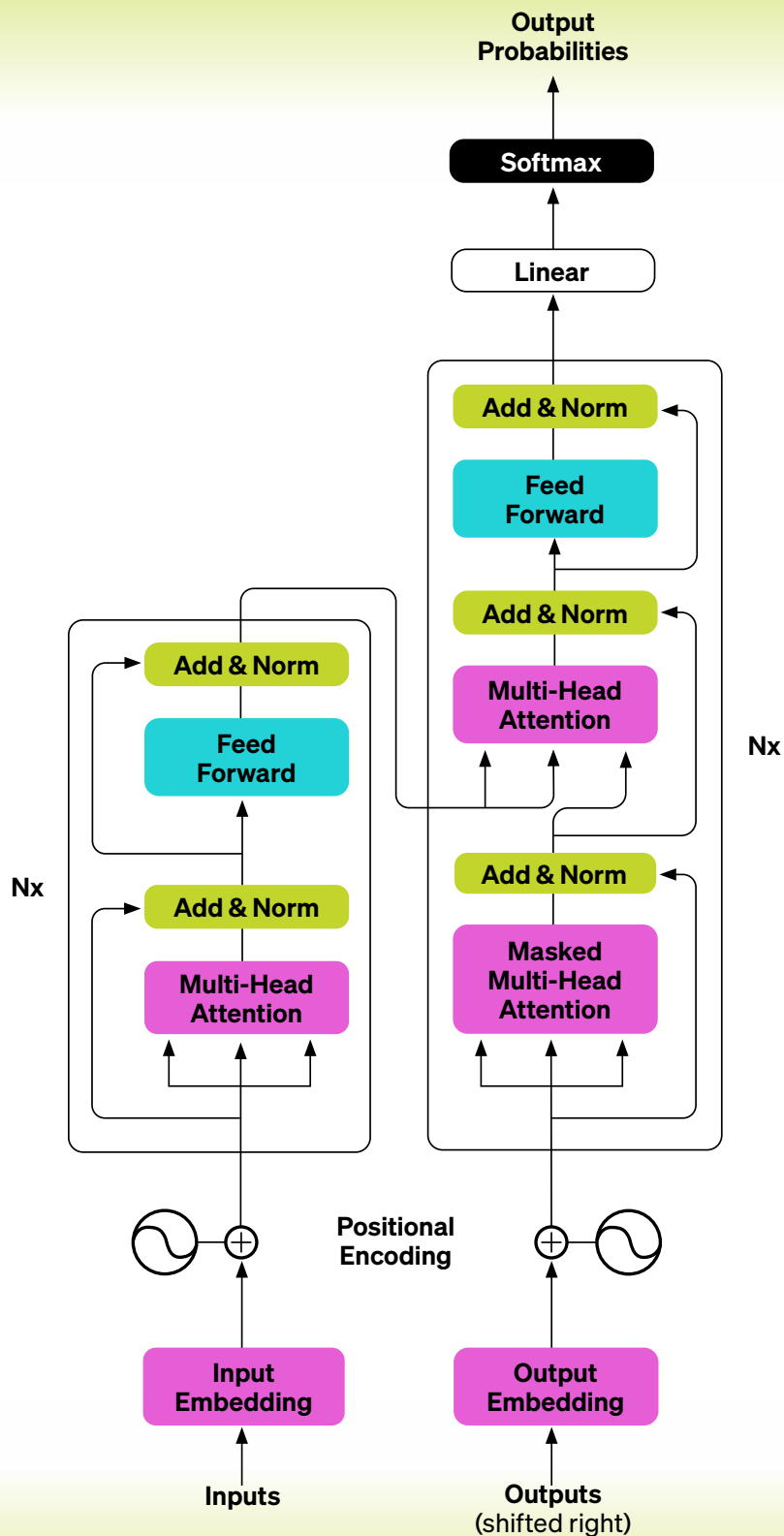


pairs; and (3) the optimization of a reinforcement learning (RL) algorithm by fine-tuning the SFT for alignment—improving the text outputs so that they are natural-sounding and safe. (See page 25 for a detailed discussion of prompt engineering, a component of SFT.)

The algorithms that power ChatGPT—LLMs and the RL algorithm of Proximal Policy Optimization (PPO)—have existed for a while. Innovative research in the field of NLP began to emerge a decade ago when neural machine translation was introduced using an attention mechanism<sup>36</sup> in RNNs.<sup>37</sup> In 2017, the attention function was a key design element in the transformer architecture introduced in the influential paper “Attention Is All You Need.”<sup>38</sup>

The transformer neural network architecture, which specializes in sequence-based inputs (such as sentences), is the backbone of LLMs (Figure 12). Multiple blocks of attention functions are integrated into the transformer architecture, along with mappings of word position and values, to weigh the importance of different parts of the input sentence. Unlike RNNs that preceded transformers and had to process data one at a time, the transformer architecture processes data in parallel. This enables it to handle longer sequences (like sentences) and context, which is why it has achieved such dramatic success in machine translation, text summarization, and image captioning.

Figure 12: The Transformer Model Architecture<sup>v</sup>



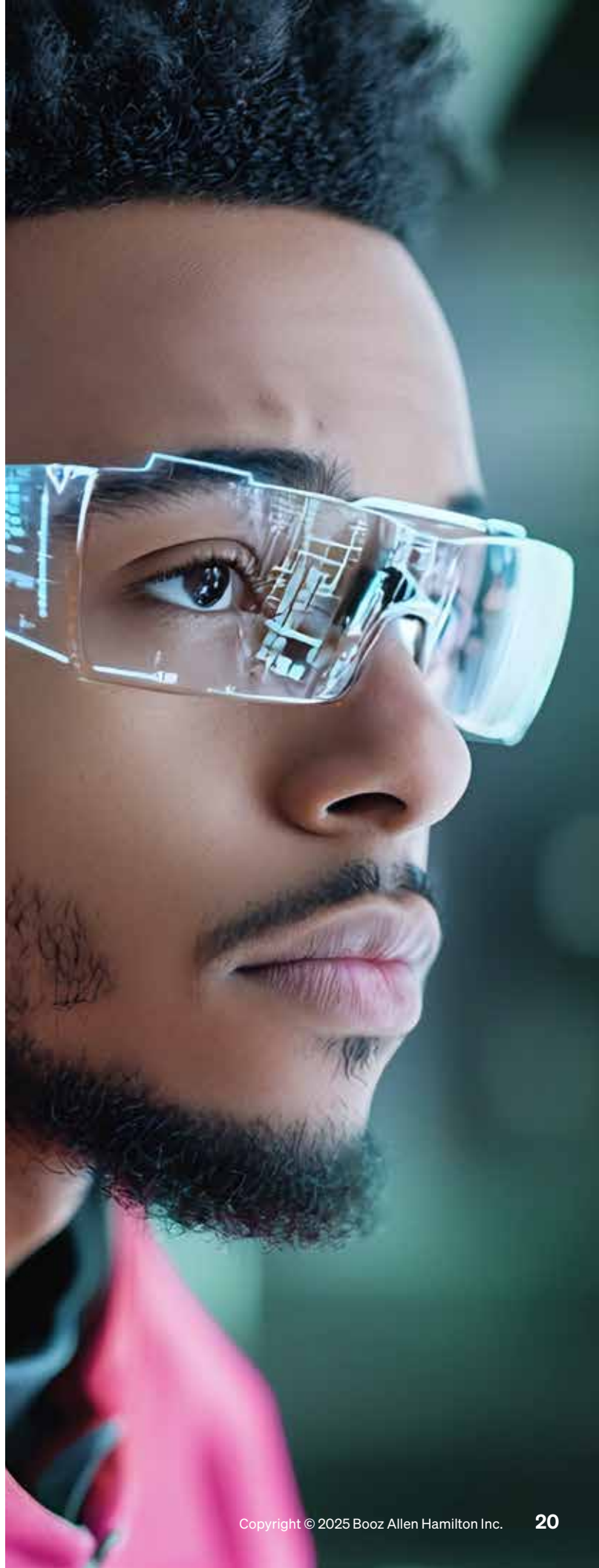
## 2022: Multimodal AI

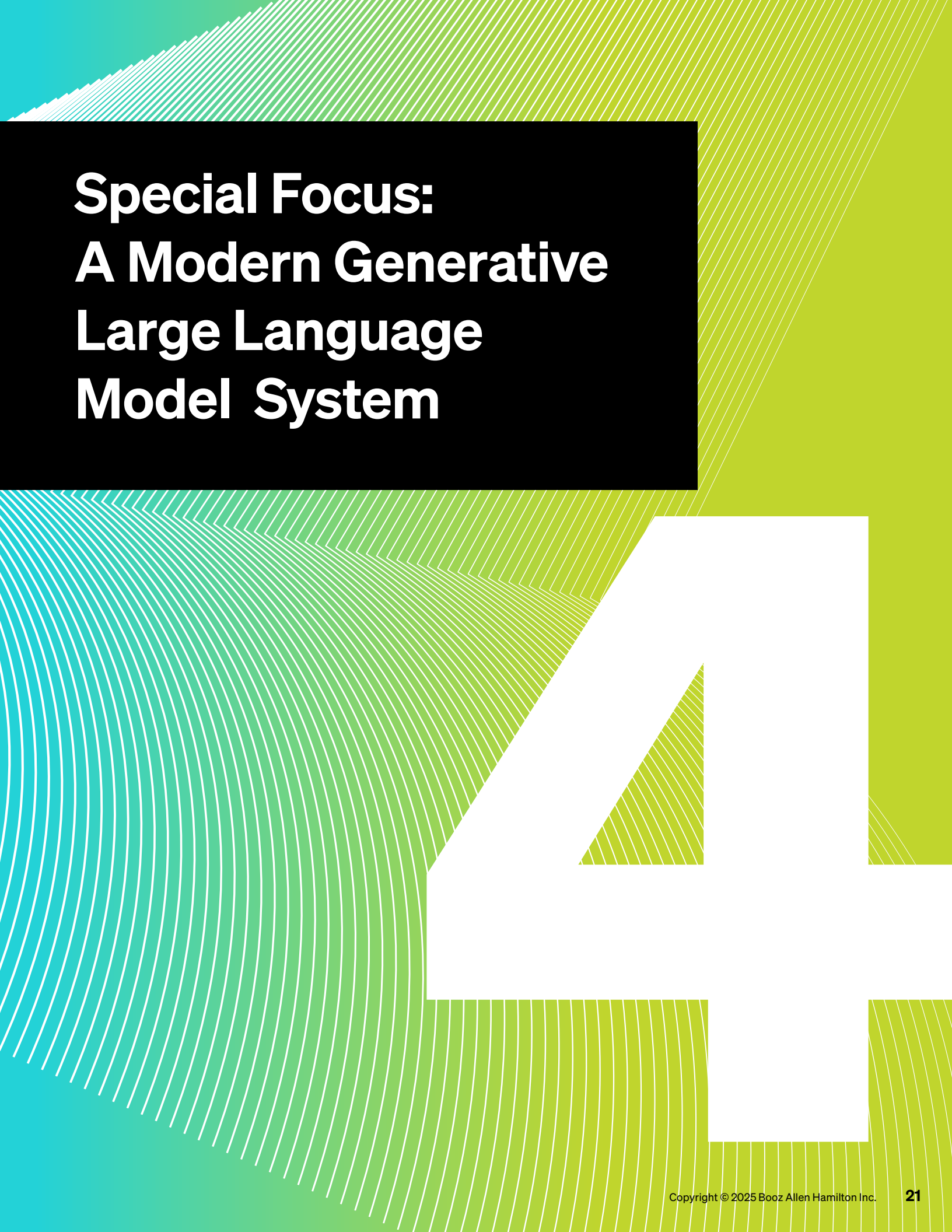
In the early 2000s, the rise of big data made waves across industry. The three v's of volume, velocity, and veracity became a tagline when it came to ML, along with the notion of “unstructured data.” To that extent, many standalone datasets for imagery (e.g., ImageNet), NLP (e.g., the popular 20 newsgroups dataset), and audio (e.g., LibriSpeech) were available for training. However, much of today's data is not only complex, voluminous, and streaming, but also multimodal,<sup>39</sup> meaning that multiple data types are necessary to train AI algorithms simultaneously. Examples of multimodal datasets include image captioning (e.g., MS COCO), question-answering (e.g., YouTube2Text), and semantic analysis (e.g., Wikimedia Commons).

The notion of multimodal data and even MMAI is not new, and as a result, it's difficult to pin down the exact date when MMAI took off. But what makes MMAI valuable are recent use cases to improve the accuracy of prediction by combining multiple signals instead of relying on just one. For example, the National Institutes of Health's National Cancer Institute is using MMAI to predict chronic cancer pain by fusing visible and thermal imagery, text, and audio, as opposed to relying solely on facial images.<sup>40</sup>

MMAI also offers generalized frameworks to solve tasks less explicitly (i.e., “zero-shot” or “few-shot” learning). Early models for affective, or sensory, computing incorporated audiovisual speech recognition. Almost all of today's AI apps for music, movies, television, retail, autonomous driving, health, food, and other areas of life rely on fusing together multiple data types—multimodalities. In this respect, a single algorithm that may consist of one neural network (such as a transformer) or multiple neural networks (such as a CNN and LSTM) and will accept multimodal data as inputs all at once.

As previously discussed, one of the most popular multimodal algorithms is OpenAI's CLIP algorithm,<sup>41</sup> which is used in the pipelines of AI art generators like DALL-E and Midjourney. As a state-of-the-art visual classifier, CLIP learns visual concepts directly from natural language. It is pre-trained on hundreds of millions of pairs using two models simultaneously. CLIP uses an image encoder like a vision transformer<sup>42</sup> to learn image features and a text encoder (a 63 million-parameter transformer) to predict the correct pairing of an image and text. Unlike traditional models for image captioning, CLIP predicts the most probable caption without being explicitly trained on image/text pairs and can therefore accomplish many computer vision tasks in a “zero-shot” fashion.





# **Special Focus: A Modern Generative Large Language Model System**

GenAI might seem almost magical as it rapidly pulls off creative and analytical feats that take far more time for human beings to accomplish. In this section, we'll unpack some of the technical capabilities that operate in concert to bring GenAI, and its profound impacts, to life. We'll focus on large language models or LLMs, which are the foundation for today's GenAI systems.

## Large Language Models

LLMs are general-purpose deep learning models that summarize, classify, and produce content. They are trained on massive amounts of data, often petabytes in size, which creates a vast number of trainable parameters (i.e., 1 billion or more) for generating

dynamic responses to prompts or questions. The process of training LLMs involves numerous steps. Seemingly every week, a new LLM is released that advances current techniques but minimizes compute requirements, boosts performance, or cleverly mixes models. The complex branching of the LLM family tree suggests the series of interrelated efforts that led to powerful models like Anthropic's Claude, Cohere, Google's Gemini, Meta's Llama, Mistral, and OpenAI's GPT-4.o (Figure 13). But what, exactly, goes into optimizing the performance of the powerful GenAI systems that have resulted from this ongoing progress? Here are a few ways we can improve LLMs with fine-tuning, agents, RL, prompt engineering, and retrieval-augmented generation (RAG).

Figure 13: The Modern Evolutionary Tree Of Llms<sup>43</sup>

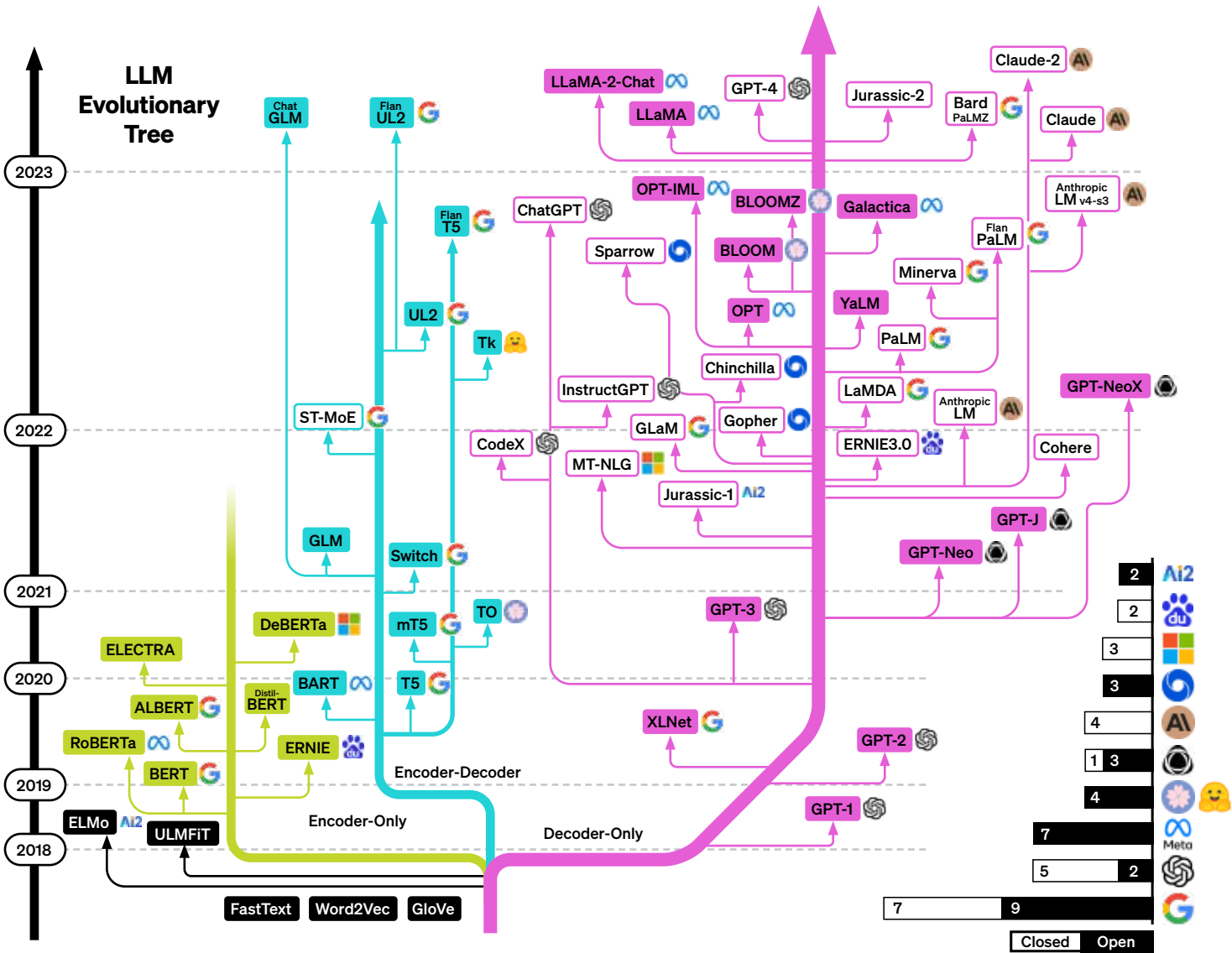
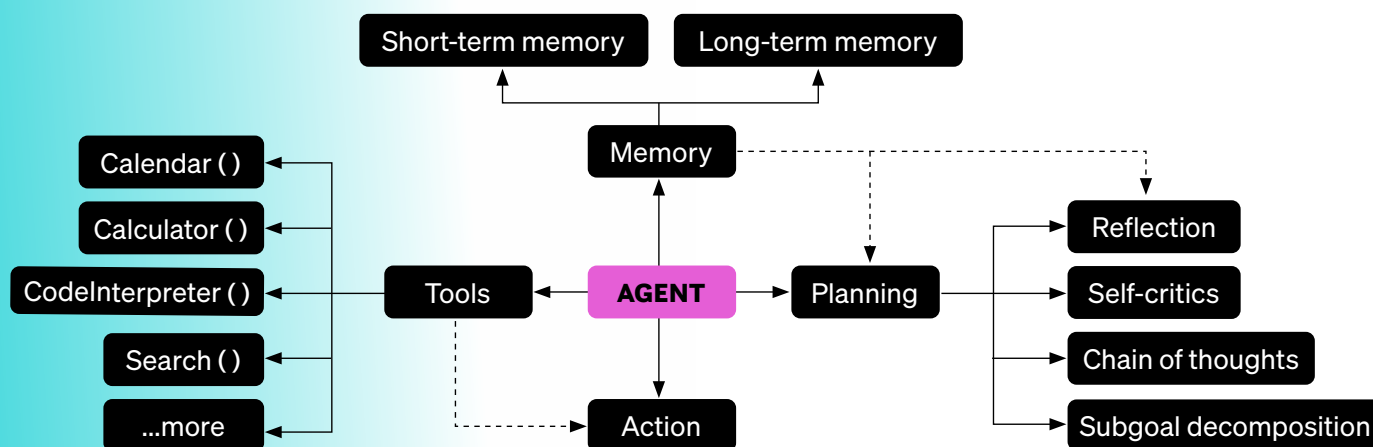




Figure 14: LIM Agents (Reference: Lilian Weng, LIM Powered Autonomous Agents)<sup>46</sup>



## Fine-Tuning

Fine-tuning originated as a transfer learning method for NLP tasks. It enables engineers to augment the knowledge of a large model (such as an LLM), which has been pre-trained on large volumes of data, with a much smaller dataset in order to customize the outputs.

Data for fine-tuning LLMs is typically offered in question-and-answer pairs that resemble an extremely minimal set of training data. The underlying data can be general in nature to approximate human knowledge or targeted to specific domains to improve performance and precision. Since the foundational LLM has been pre-trained on an entire corpus of semantics and syntax, it can infer patterns based on the minimal samples. Here the LLM retains most of the knowledge it gained through pre-training (these parts of the model are “frozen”)—but the rest of the LLM updates with new information.

Fine-tuning a foundation model involves updating some, but not all, of its weights to better perform a specific task. However, if overdone, this process can lead to “catastrophic forgetting,” where the model loses its ability to perform well on tasks it was previously good at. This is a potential risk for highly complex government mission applications for the U.S. national security community or DOD, for example, as they adapt commercial models to some of their highly specialized requirements. Balancing fine-tuning requires adjusting just enough to improve performance on a new task without erasing the original knowledge to ensure the model remains versatile and effective across various tasks.

One effective fine-tuning approach is using low rank adaptation (LoRA), which was introduced by Microsoft as a parameter-efficient process.<sup>44</sup> The key idea with LoRA is to store and load fewer parameters (e.g., “weights”) for each fine-tuned model, while keeping the pre-trained base LLM model weights frozen (i.e., unchanged). Since fine-tuning for a specific task—for

example, learning the personality of an innkeeper in *Dungeons & Dragons*—is fairly narrow, the full set of LLM weights does not need to be completely updated, thus improving model performance without undermining proficiency in earlier tasks.

## Agents

With increased capability over previous ML models, LLMs, in concert with other software, may be able to assume the role of an independent actor (or agent) within a larger system. Deploying LLMs in this configuration could partially automate operations or one day give rise to fully autonomous systems. Agent-based systems enable this.<sup>45</sup>

Agents could perform various tasks when automated, including planning and executing, applying memory, and using tools (e.g., opening a website; formatting text) (Figure 14). The figure shows the four components the agent leverages: tools, memory, planning, and action. In a traditional setup, the agent will have access to non-LLM software such as application programming interfaces that can open calendars or search the internet. The use of these tools is triggered by a set of actions that prompt the agent to call them.

The agent uses planning to do more with the prompts by reasoning through and evaluating a series of steps or outcomes. This might include decomposing a larger, broader prompt into a series of smaller goals where each output is then a new answer or result. Memory comes in the form of vectors where the prompt is tied to a series of outcomes such as “false” or “true” to store history about the success of tried attempts. The combination of agent planning and memory is an exciting area of engineering that is enabling LLMs to be used for more than chat or summarization—such as imitating the reasoning that occurs from training challenging RL algorithms.

To provide an agent with operating context, engineers program the agent with a set of constraints (i.e., rules) it must adhere to, commands (i.e., tools) it might use (e.g., retrieving a spreadsheet), and sub-tasks it must complete (e.g., itemization, writing, and descriptive analysis). LLMs can complete difficult tasks and vary significantly from traditional “bots” such as non-AI tools like robotic process automation software (Figure 15).

Robotic Process Automation	Large Language Models
<ul style="list-style-type: none"> <li>Designed to automate specific, pre-defined tasks in business workflows.</li> <li>Rule-driven, executing only the precise steps that programmers have explicitly outlined (e.g., entering data).</li> <li>No ability to learn or adapt independently.</li> </ul>	<ul style="list-style-type: none"> <li>Designed to orchestrate a chain of actions in concert with other agents.</li> <li>Broadly responsive to a series of simple user prompts, performing an array of often complex tasks (e.g., crafting analyses).</li> <li>Capacity to learn how to reason.</li> </ul>

Figure 15: Comparing Robotic Process Automation With Large Language Models

## Reinforcement Learning With Human Feedback

Essentially, RL is an area of ML where agents learn behavior patterns via trial-and-error interactions with a dynamic environment (Figure 16). RL has famously been used to train algorithms to play games at levels far beyond human ability, as well as to execute robotics, and increasingly, to align preferences for GenAI output.

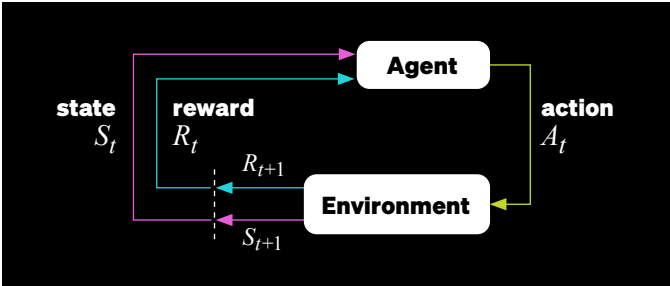


Figure 16: A Traditional Reinforcement Learning Framework<sup>51</sup>

RL has a long history, and it emerged out of established, multidisciplinary methods in computer science, psychology, control theory, and statistics.<sup>47</sup> The fundamental principle is that an agent will connect to an external environment. As the agent interacts with the environment, it will take an action.

What will the agent ultimately be designed to do? If the environment were a game of chess, for example, the agent would scan the board for all possible moves. Then it would need to take action to achieve a goal, like

putting the opponent’s king in check. After that it would receive feedback—a “reward” or a penalty, such as losing a piece on the board. The more it plays, the more it learns about how to play. One advantage of RL models is the way data is collected for training—it gathers data dynamically through these interactions with the environment.

RLHF is another application of RL, but this time for improving the alignment of generative LLMs as detailed in the InstructGPT paper.<sup>48</sup> RLHF is a key component of state-of-the-art generative LLMs. It consists of three basic steps:<sup>49</sup>

- 1 Researchers create a database of prompts and ask humans to manually write responses to each prompt, creating a rich set of instruction-output pairs. This dataset can then be used to fine-tune a pre-trained base model (e.g., GPT-3) in a supervised fashion (i.e., with an SFT).
- 2 For each prompt, the SFT outputs several responses. A human ranks the responses based on their preference—and, with that, we now have a dataset of prompt, response, and associated preferences. This dataset informs the design of an RM algorithm that accepts a sequence of text and returns a numerical reward that represents human preference.
- 3 The Proximal Policy Optimization (PPO) model<sup>50</sup> uses RL and the reward scores outputted by the RM from its regression layer to fine-tune the SFT. Today a variety of RL models are used beyond PPO. But in the traditional InstructGPT framework, PPO has been shown as effective.



When there are so many RL algorithm choices, why is a PPO model so useful? <sup>52</sup> PPO offers greater training stability by eliminating the need for large policy updates. Because the PPO algorithm enforces a series of constraints that limit how large the learning steps can be, it updates the policy using smaller steps, which is likely to converge to a more optimal solution. The idea is to control the learning to avoid updating the RL model's weights in a destructive way.

## Prompt Engineering

While anyone with an internet connection and email address can use ChatGPT and other commercial LLMs immediately, skilled users who understand how the underlying model works will extract significantly more value from this technology. This is because expert prompt engineering helps steer the LLM to produce outputs that users can, with confidence, directly integrate into codebases, databases, and various components of a system. Better prompting improves responses in image generation as well as text generation and anywhere prompts generate content (Figure 17).

*Figure 17: AI Art Generators Vary, But Effective Prompting Always Improves Images*



**Midjourney With A Poor Prompt**  
"A Delta IV heavy rocket"



**Midjourney With A Good Prompt**  
"I'm a NASA photographer. I just took a picture of a Boeing Delta IV heavy rocket lifting off from the launch pad at Cape Canaveral on a beautiful sunny day."

**There are four basic principles for prompts that every user should adhere to:**

- 1. Show and tell—be very clear in the instructions.
- 2. Break work into smaller, discrete chunks.
- 3. Provide high-quality data—give good examples and proofread prompt instructions.
- 4. Prompt the LLM to check and improve its own output.

**Considering a few additional guidelines can further improve results:**

- + Explicitly state the desired focus, format, style, intended audience, and text length.
- + Create a list of topics or points to cover.
- + State the perspective from which the text should be written.
- + Specify requirements, such as “Add as many quantifiable references as possible.”
- + Break up prompts for long-form content into small pieces.
- + Adopt a coder’s mindset of “programming a machine,” not “conversing with a human.”
- + Use cheat phrases for efficiency (Figure 18).

Prompt Template	Purpose
“As a [insert profession/role]”	Frames the LLM’s knowledge.
“In the style of [insert famous person’s name]”	Enables style matching.
“Explain this topic for [insert specific age group or job].”	Defines the audience.
“For the [insert company/brand publication]”	Adjusts voice and tone.
“Let’s think step by step.”	Elicits a process description.
“Why do you think that?”	Asks the LLM to explain its reasoning.
“Thinking backwards”	Helps the LLM retrace its steps when it keeps outputting incorrect conclusions.
Negative prompts: “Do not use numbered lists.” “Avoid acronyms.”	Tells the LLM what to exclude from its response.
Specify references: “Include only a reference that is widely cited in the literature.”	Improves accuracy; avoids the invention of fake references.

Figure 18: Some Useful LLM Cheat Phrases



# Retrieval-Augmented Generation

A final way to enhance standard language model responses is by using RAG to incorporate external, relevant information into the generation process.

Unlike traditional models that generate responses based solely on their training data, RAG leverages a database search to enrich the response to a user's prompt. In contrast to standard language model generation, where an LLM responds to a prompt based on the content on which it was trained, RAG exposes the LLM to additional, context-relevant information that can be used alongside a prompt. The LLM accesses this information by searching a database of content using the user's initial prompt. This happens with semantic search, which is the way most search engines find content, or through a vector search.

## How Vector Search Works

A component of the broader semantic search process, vector search is a method for finding text that is numerically similar to a user's prompt. Instead of focusing on keywords, this method finds similar or related concepts that would not normally appear as synonyms. In vector search:

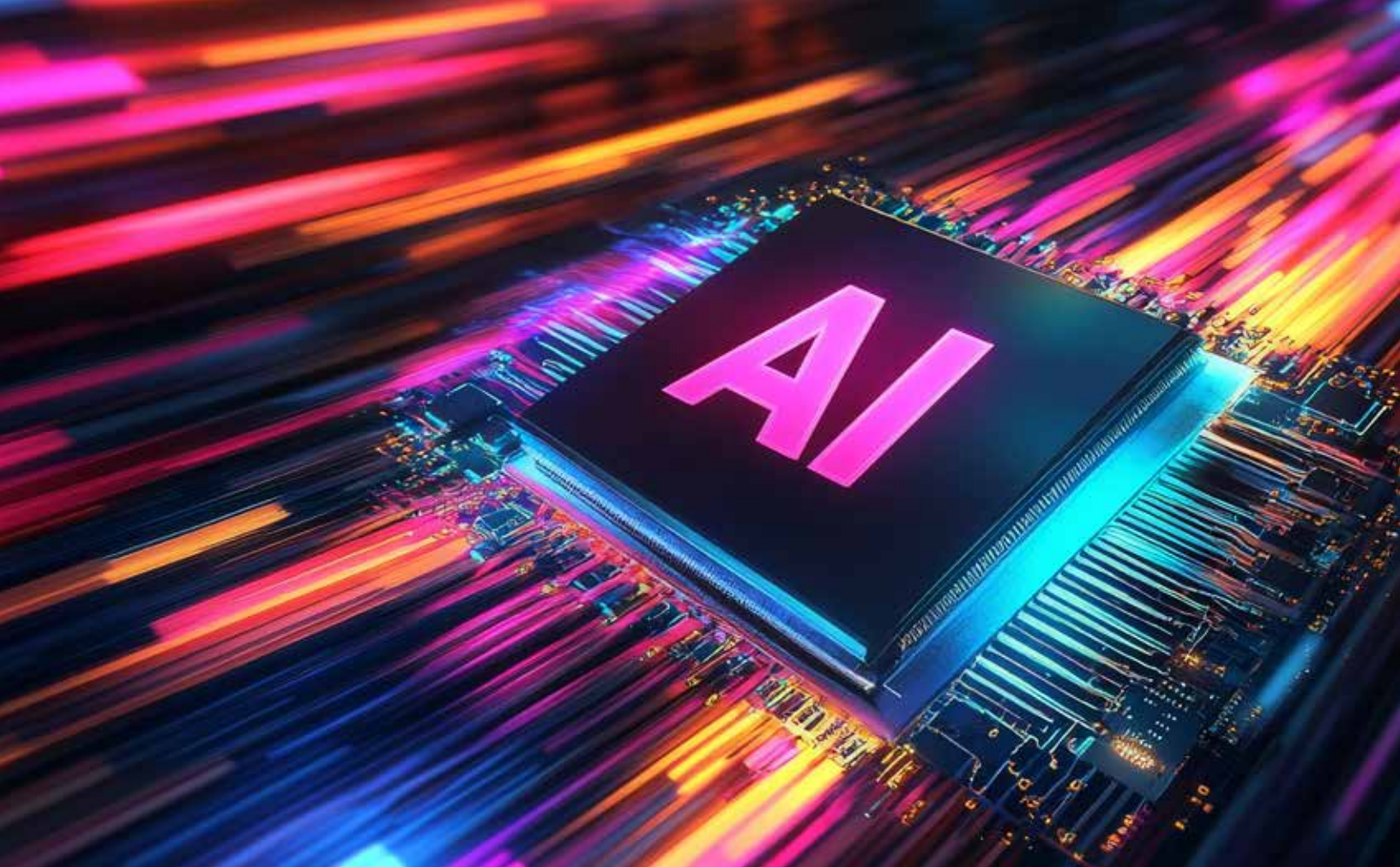
- Chunks of text are compressed into fixed-sized lists of integers known as “embeddings.”
- The user prompt is also “vectorized” or “embedded” into a similar list of integers.
- The two embeddings can then be compared to see how close their meanings are.
- Through these comparisons, vector search returns the most similar piece of content.

Semantic searching is more complex and requires more computation to determine the intent and context behind text. As a result, vector searches are often preferable to semantic searches when time or compute resources are limited.

RAG is particularly important when an LLM must be factually correct or consistent with a corpus of information on which the foundational model was not trained. For example, a model used to make healthcare recommendations could be improved by granting it access to medical information, such as common procedures, symptoms, or given appropriate data protection, the patient's medical history. Providing the model with factual content relevant to the user's request can reduce the likelihood of a hallucination (see the “Generative Hallucinations” section of this primer), not unlike how humans can easily make mistakes when relying entirely on their memory but can avoid those mistakes by referencing written material.







## Accelerated Computing

### Hardware Innovations That Power AI

- Accelerated computing goes hand in hand with AI. Harnessing complex AI models requires a basic understanding of how hardware such as GPUs is leveraged not only to train AI models but also to run prediction and inference. While some traditional ML algorithms (such as support vector machines) do not require specialized hardware, the modern AI algorithms discussed in this primer require training and inference on a GPU—and often several.
  - A GPU is a specialized processor that breaks up tasks and runs them in parallel using thousands of cores to spread out mathematical operations for tasks like deep learning and graphics rendering. A NVIDIA A100 GPU, for example, consists of 6,912 cores, and its RTX 2080 Ti GPU consists of 4,532 cores.
  - Today's image classification models can run on a single GPU with as few as 11 gigabytes of video RAM (VRAM), which is similar to the RTX 2080 Ti.
- However, algorithms such as traditional transformers are computationally expensive to train and require more VRAM and capacity.
- When algorithms are computationally expensive, understanding GPU speed is essential. One way to measure computational speed is through the floating-point operations per second (FLOPS) metric. While the RTX 2080 Ti performs at approximately 13 terraFLOPS, an A100 delivers 312 terraFLOPS.
  - A high volume of computational operations can unacceptably slow training times. For example, training GPT-3 (175 billion parameters) on 8 NVIDIA V100 GPUs would take 36 years. Compare this to training it in 7 months using 512 V100s.<sup>53</sup>
  - As AI models become increasingly expensive due to their size and capacity to learn complex data, GPU manufacturers are releasing processors capable of scaling parallelism and decreasing latency, while AI software is also being offered to manage the neural network itself.

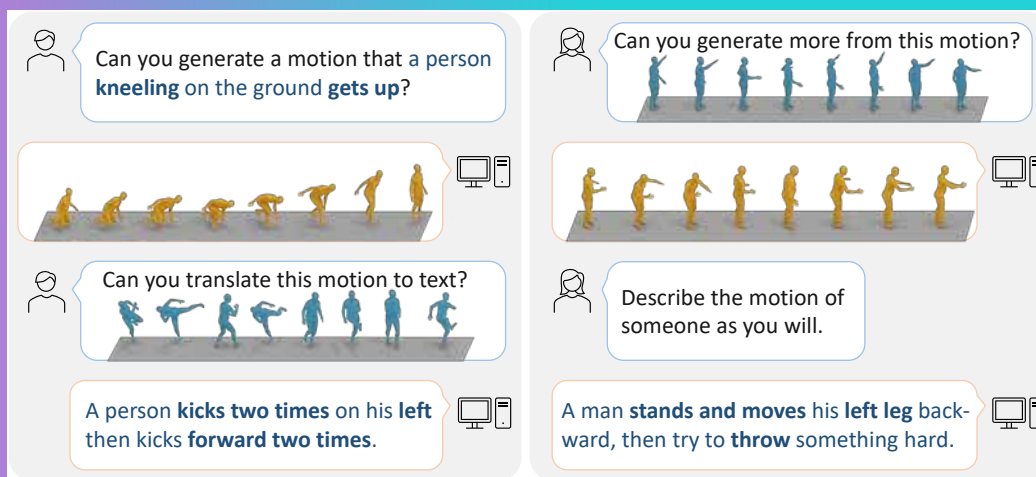
## When To Use GenAI

Organizations can use GenAI to operationalize any number of applications, from customer service chatbots and software coding agents to tools for scientific discovery and policy adjudication. Many of these use cases are supported by GenAI's ability to provide automated knowledge and data management encompassing a series of integrated functions:

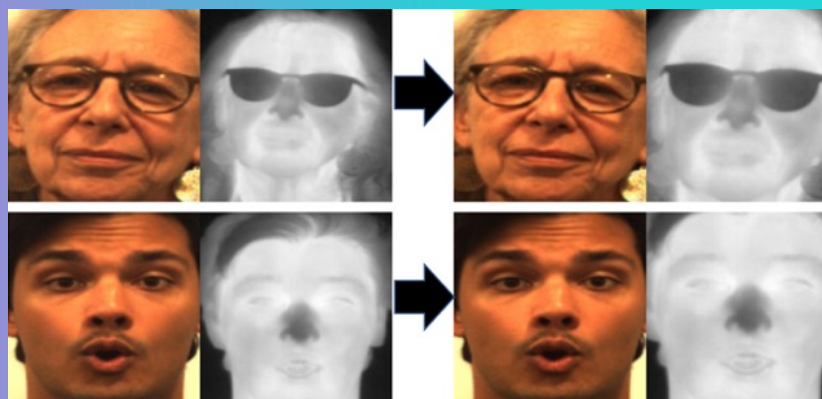
- Search, aggregation, and summation
- Interpretation, analysis, and synthesis
- Generation, prediction, and interaction
- Iterative design and development

Leveraging these capabilities creatively can unlock a host of new possibilities, as we are witnessing today. When we use them as a computational tool, GenAI algorithms offer a sophisticated means of unsupervised learning (i.e., GenAI methods generate images and text not from reference samples but from learned representations of data).

Yet GenAI encompasses more than applications like ChatGPT that may be most useful for purely generative outputs. For example, in a lab setting, researchers can use GenAI algorithms to generate a code bank of labels that represent human motions (Figure 19). Encoding the action of a person kneeling on the ground and standing in MotionGPT<sup>54</sup> takes much less time than labeling by hand every step in this process (e.g., kneel, kneel higher, slightly squat). Likewise, VistaMorph<sup>55</sup> is an example of two GANs working in a pipeline to calculate warp, zoom, and perspective when the actual coordinates are not available, especially across millions of images.



MotionGPT—GPT to generate human motion



VistaMorph—GANs to learn how to align images

However, with all the media hype GenAI has received, it can be difficult to remember that GenAI is merely a tool—albeit an advanced one—that occupies a new space in the arsenal of AI algorithms. While GenAI can be applied to a variety of challenges, it cannot solve every imaginable problem. For this reason, it’s best to avoid thinking of it as the solution for all cases, especially considering the continually expanding utility of traditional AI (Figure 20).

	Traditional AI	Generative AI
<b>Data Requirements.</b> Volume and quality of data needed for accuracy.	● ● ● ●	● ● ● ● ●
<b>Expertise Needed.</b> Level of specialized knowledge required to use effectively.	● ● ●	● ● ● ● ●
<b>Computational Costs.</b> Resources and expenses for processing and running models.	● ●	● ● ● ● ●
<b>Time to Deployment.</b> Duration from development to operational use.	● ● ●	● ● ● ●
<b>Interpretability.</b> Ease of understanding model decisions and outputs.	● ● ●	● ● ● ● ●
<b>Output Accuracy.</b> Precision and reliability of the generated results.	● ●	● ● ●
<b>Output Creativity/Reasoning.</b> Ability to generate novel ideas and logical conclusions.	● ● ● ● ●	●
<b>Output Controllability.</b> Degree of influence over generated results.	●	● ● ●
<b>Generalizability.</b> Ability to apply learned knowledge to new, unseen data.	● ● ● ●	●

- Well-suited application of the technology with a low threshold to implementation
- ● ● ● ● Less optimal use of the technology with a high threshold to implementation

Figure 20: When To Consider GenAI For A Use Case



## Creativity Versus Certainty

In general, use cases that require definitive answers such as medical diagnoses, network security monitoring, or supply chain optimization may be better suited to traditional AI solutions. In contrast to traditional AI models, which are deterministic, GenAI models are, by design, stochastic, delivering outputs that are probabilistic in nature and often resulting in some randomness or variability in their responses. This GenAI characteristic is why, given the same exact input or prompt, the output will vary even when users are looking for a single, authoritative result. Use-case analyses should factor in this critical difference to allow organizations to accurately calibrate operational and mission risks. For example, traditional AI is a better—and less risky—choice for use cases where the problem is well-defined and has specific rules and boundaries, like automating quality control in manufacturing using image recognition or detecting fraud across financial institutions. Assessing the tradeoffs across the available AI types will help organizations identify whether GenAI or other forms of modern AI offer the better approach.

## AI Operations

### Keys To Seamless Deployment And Scaling

As models grow more complex, different components and pipelines must be blended in a manner that scales and accounts for changes that may cause downstream impacts over time. Collaboration across teams results in enterprise-ready deployments faster, with less overall effort, and with lower operations and maintenance costs that drive down the total cost of ownership.

- Fielding enterprise-class AI systems requires the application of modern software engineering practices for continuous integration and continuous delivery to facilitate reliable, high-quality deployment of AI models.
- However, AI systems differ from standard software delivery efforts because of their inherently recursive nature. In accounting for this difference, data scientists leverage an exploratory-analysis process to understand the data available, experiment with appropriate data preprocessing, and train/test models.
- The data scientists and AI engineers then work together to harden science artifacts via software engineering. Specifically, data preprocessing routines are extracted into modular, reusable software components that take advantage of modern programming techniques. These techniques include parallelization of processing and the use of DevSecOps to achieve repeatability, testability, and scalability.
- As stakeholders request more standards for governance and transparency, toolkits to check off compliance with ethical AI guidelines are needed.
- Finally, system tools to scan for malware and assess models' security fitness help thwart malicious actors that seek to poison and compromise AI models early in the process.



# Recognizing AI's Limits And Challenges





GenAI's strength is its ability to respond dynamically with distinct replies to each unique prompt along with its potential to learn and improve. Not surprisingly, these strengths also introduce constraints and challenges that can impact the accuracy, relevance, and integrity of these responses. What special GenAI-related risks should organizations be aware of?

## AI Bias

For several years, critics have scrutinized AI for its vulnerability to bias that causes models to output “unfair” decisions. Since most Western AI models, to date, share the same datasets (e.g., ImageNet, MSCOCO, CelebFaces, Wikipedia, Books Corpus), any such biases could become entrenched in production systems worldwide. While there is no standard definition of “fairness” in AI, a useful shorthand is “the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.”<sup>59</sup> Occurring as a result of the AI algorithm, user interaction with the model and the data used for training bias can drive unfair AI outcomes.

Ideally, an AI tool would not tend toward any human, systematic, or institutional bias—but it is not possible to entirely eliminate such bias. Given this reality, it is the responsibility of each organization to recognize inherent

bias, mitigate as much as possible, and ultimately provide transparency to users and stakeholders (Figure 21). The latter point is especially important, given the prevalence of bias embedded across society, as it allows users to make conscious decisions about how to interpret the data.

There are several forms of bias, including:

- **Measurement bias** that comes from how certain features are used as mismeasured proxy indicators. An infamous case is using prior arrests in the COMPAS prediction tool as a measure of “riskiness.”
- **Simpson’s Paradox**, a statistical phenomenon where a trend can appear, reverse, or disappear based on combining groups of data.
- **Algorithmic bias**, where bias is injected by the algorithm as a result of the model’s architecture and training parameters.
- **Historical bias** from sociotechnical issues that seep into data generation even prior to sampling, such as search results revealing fewer women chief executive officers than men due to less representation in the workforce during a specific time.
- **Population bias** amplifying demographic differences in a user population (e.g., more women use Pinterest than men, whereas men are more apt to use X/ Twitter).



While today’s data scientists and ML engineers draw upon significant experience in addressing and mitigating bias, the rapid pace of AI evolution and the expanding scale of models create unprecedented challenges. Enterprises can take steps to address bias by ensuring the use of a diverse and robust dataset. It is also crucial for organizations to involve key stakeholders throughout the lifecycle of an AI tool, from conceptualization through its eventual

decommissioning. Beginning with the initial business use case and advancing through substantive development, decision making, and continuous monitoring, stakeholders should have a clear understanding of the AI’s function and role at each step. This approach facilitates informed decision making and enhances transparency and accountability to monitor for and address any underlying biases in the AI’s output.





 <b>Bias Scenario</b>	 <b>Potential Risks</b>	 <b>Key Questions</b>	 <b>Mitigation Strategy</b>
Datasets that span several years may be outdated.	Bias can unduly impact older data due to poor processes, manual input methods, poorly documented changes to applications and workflows, and lack of annotation of the associated datasets.	<p>What data elements are planned for collection, and how are they essential for the system’s functionality?</p> <p>How does the system design ensure that there is no excessive data collection?</p>	Pair data scientists with business process experts to create AI R&D teams that work to understand the data sources, relationships, meaning, and impact of business process changes before presenting results.
New/unidentified sources of bias may emerge with ongoing use.	The risk of encountering new or previously unidentified biases increases as AI applications scale and process larger volumes of data over time.	What biases could potentially emerge from subtle correlations or patterns in the data that were not apparent during initial training?	Continuously monitor and reevaluate model performance and decision-making processes, include metrics to determine impacts to equality from the beginning, and prepare for those measures to change as the AI program evolves.
A model’s inherent prejudicial assumptions can yield bias.	This can result in inequitable outcomes regardless of the nature of the training dataset.	<p>What methods will the model use to detect and rectify biases, particularly those that arise from data inputs or algorithmic structure?</p> <p>What approaches will the model use to ensure fairness in its outcomes?</p>	Design and continuously evaluate a stakeholder-agreed process to integrate equity into the AI model, considering the full range of the model’s effects on various societal groups.
Bias or inequality can emerge as a result of failing to properly ensure transparency regarding AI usage.	Users may unintentionally introduce bias into AI-generated products if they don’t understand the data sources, types, and usage patterns.	<p>How are user consent and notice mechanisms described in the system design?</p> <p>What measures are planned for ensuring transparency in data usage?</p>	Ensure that data produced within AI models is labeled as such. Disclose the use of AI when disseminating AI-generated material.

Figure 21: Representative Bias Risks And Mitigations



# Generative Hallucinations

When AI models produce coherent grammar, but nonsensical content, we tend to call this a “hallucination.” It’s an anthropomorphic interpretation of the algorithm—that it must have “hallucinated” facts, references, names, and dates that just are not real and woven them into a grammatical sentence. Yet from the AI’s perspective, there is no difference between a hallucination and an accurately cited, factual statement.

This is because LLMs are trained to be autoregressive, meaning they predict the next likely word from the previously seen words. This training objective makes no distinction between fact and fiction—provided the next word looks plausible, it is all the same autoregressive goal. It is true that factual data, having been part of the training of the algorithm, is more likely to look correct, but the AI feels no constraint to faithfully reproduce any particular fact.

AI hallucinations are a natural consequence of how GenAI works today. During the training of generative imagery models such as GANs and diffusion models, hallucinations, and distortions—such as stippling patterns, checkerboard effects, and/or low diversity of images—often creep into the learning process. In general, there are two forms of mislearning that occur within GenAI models: mode collapse and model collapse.

## MODE COLLAPSE

Specifically in GANs, hallucinations can be caused by “mode collapse,” a term introduced in 2013, which occurs when the model fails to converge, meaning it can no longer learn.<sup>60</sup> The outcome of mode collapse is the generation of images that have very low diversity, meaning that the bulk of generated images tend to favor one kind of style or object, as opposed to the variety of others it was trained on.

Why does this happen? As with all generative models, the GAN essentially plays a game of moving data distributions between what is real and what is fake. This game is constrained by complex optimization formulas that negotiate learning between the generator and the discriminator (see the “Advances in Modern AI” section for more information). Essentially, in this game, the generator will cheat and discover that there are certain distributions (“modes”) that fool the discriminator time and time again. This “mode seeking”<sup>61,62</sup> behavior compels the generator to rely on non-diverse, safe modes so that the discriminator will not penalize it.

Unlike GANs, diffusion models do not have a generator or discriminator. As a result, there are other mechanisms by which hallucinations arise. One source is the variation in the denoising process where the fake image is being generated. If errors start early during training, they will accumulate at every step of the long

diffusion process. Evidence of hallucinated artifacts can be seen with low resolution, contrast, boundary artifact, and insufficient image reconstruction issues.<sup>63</sup>

## MODEL COLLAPSE

Scraped data from the Internet has been the primary source of GenAI training data. It is likely that most original human-crafted content has already been exhausted in the training of existing LLMs. As a result, LLMs can suffer what is called “model collapse”<sup>64</sup> where the LLM degrades and becomes useless (Figure 22). The result of retraining on generated, synthetic data will lead to outputs that are non-sensical and hallucinatory. This can occur in LLMs and diffusion models for imagery. Introduced in 2023 in “The Curse of Recursion” article, model collapse is defined as “a degenerative learning process where models start forgetting improbable events over time, as the model becomes poisoned with its own projection of reality.”

The challenge will be the scale needed to make a meaningful improvement to the LLM in curating human data. One opportunity is for users to apply AI in an augmenting way (i.e., users add their own tweaks, edits, and modifications to an AI’s output). This may help reduce the model collapse issue, but the evolution of how we share data and communicate is still to be determined in the wake of LLMs.

Methods for fixing hallucinations after the training is done are now being researched and are showing some initial promise. The most popular technique is to combine GenAI with a search engine and answer questions based on retrieved documents. If users include only the documents they know to be true, they will get more reliable answers. But integrating the concept of knowledge and truthfulness is still an open and challenging problem.

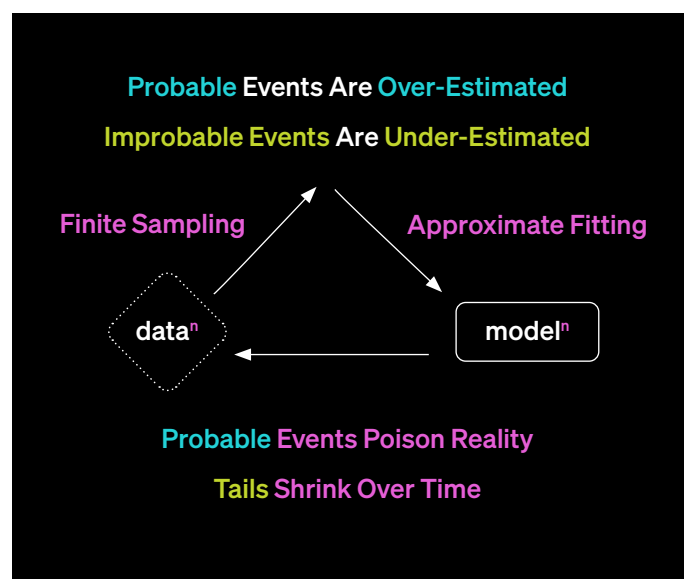


Figure 22: Model Collapse By Shumailov Et Al.<sup>65</sup>



## AI Vulnerability

As AI systems become progressively more central to defense, intelligence, civil, and commercial operations, the need to secure those systems also becomes increasingly important. Damaging attacks against AI systems—from manipulations of the physical world to fool self-driving cars to poisoned training data that compels LLMs to produce undesirable output—are being launched on commercial and government entities by adversaries.

The AI security community has responded with a variety of resources aimed at defining a vocabulary for discussing AI vulnerabilities, understanding common adversarial techniques, and even prioritizing top AI security concerns. In broad terms, adversaries employ the following five types of attacks to debase, evade, and exfiltrate predictions and private information used in AI systems:

- **Poisoning:** Adversaries pollute training data such that the model learns decision rules that further the attackers' goals. This is possible by altering only a small fraction of the training data, and it represents a growing threat given the increased popularity of foundation models pre-trained on data scraped from the web.
- **Malware:** Adversaries package malware within models such that the malware is executed when the model is loaded or when a particular node in a neural network is activated. In addition, common ML libraries have been compromised with malicious dependencies.
- **Evasion:** Adversaries engineer inputs with manipulations that result in the model making misclassified and/or unintended predictions. If not caught, these errors can result in dangerous behavior of downstream systems. Adversaries can often make evasive maneuvers with little cost; for example, inexpensive adversarial stickers/patches can fool a state-of-the-art computer vision model and low- or no-cost prompt injections can force LLMs to relinquish their guardrails.
- **Inversion:** Adversaries exfiltrate private or revealing information concerning the AI model and its training data. This can be part of a reconnaissance effort for an adversary planning a future attack or a direct attempt to seize sensitive information.
- **Model Theft:** Adversaries steal intellectual property by exactly or approximately reproducing a model. Adversaries can identify additional vulnerabilities to exploit by examining the replicated model.

Enterprises can mitigate these risks by introducing safeguards, including adversarial training, model and dependency scanning, and evasion attack detection, throughout the AI operations lifecycle.



# **A Look Ahead At Modern AI's Potential Realized**



While GenAI's technological capabilities have wowed global audiences, its enterprise impact has been more nuanced. In simplest terms, this technology has yet to consistently and independently perform complex operations with high confidence in many instances.

However, this will likely change over the next 18 to 36 months as the underlying technology matures. Most significantly, we can expect continued improvements in contextual understanding, reasoning abilities, causal inference, knowledge retention, and abstraction. Furthermore, increased integration of stochastic and deterministic approaches—modern AI—will enable more well-rounded and balanced problem solving.

We should also expect continued advancements in computational hardware that balances speed, model size, and complexity as more use cases begin to rely on large GenAI algorithms.

This progress will enable AI to tackle more complex problems, make nuanced judgments, and provide more reliable professional recommendations. As these capabilities mature, generative AI will move closer to emulating human-like cognitive processes, potentially leading to more autonomous decision making in certain domains where it has accumulated a wealth of knowledge and has been tuned accordingly.

Furthermore, it is the combination of AI algorithms that will be powerful—not a single model alone. In the past several years we have seen progress in discrete, individual neural networks. In the future, AI tools such as an LLMs combined with RL algorithms, with a graph

convolutional neural network will be combined to solve different tasks in concert. In this field, we already see research into breaking up complex systems using “estimators,” “evaluators,” and “critics” that solve problems collaboratively.

Today's GenAI can be primarily defined as a collection of tools that can summarize knowledge bases and generate probabilistic responses. However, we will soon see an increasing shift to autonomous agents that rely on reasoning to navigate and negotiate more complex operations independently.

This evolution is set to challenge enterprises in two significant ways. First, they must make urgent investments in training programs to enhance employees' AI literacy, prompt engineering skills, and AI-human collaboration techniques. While some roles may be automated—new positions focused on AI oversight, ethical considerations, and strategic application of AI technologies are likely to emerge. Simultaneously, they must also reaffirm their commitment to ethical, responsible AI, as these principles must be ingrained into these increasingly autonomous systems—providing guardrails to constrain GenAI within these agreed AI governance frameworks.

Despite AI's increasing potential, human judgment, creativity, and ethical considerations will remain indispensable in many organizational contexts. The most successful organizations will strike the right balance between AI automation and human expertise, leveraging the strengths of both to drive innovation and efficiency.



# The Path To Artificial General Intelligence (AGI)

Will AI become sentient? A theme of this primer has been that AI dramatically changed when OpenAI unveiled ChatGPT. Although earlier GPT versions existed, ChatGPT's release in November 2022 brought the idea of AGI<sup>65</sup> more broadly into public consciousness.

The definitions of AGI often vary by source—for example, should it simply match or actually exceed conventional human intelligence? But a common focus is a high degree of cognitive thinking and reasoning that would enable open-ended problem solving and decision making. This is a fundamental question that we are all grappling with today.

AGI and the related concept of “the singularity” were first popularized in 1983 by mathematician Vernor Vinge.<sup>66</sup> Futurist Ray Kurzweil explored these concepts more deeply in his 2005 book, *The Singularity Is Near*.<sup>67</sup> Both researchers characterize the singularity as an ultra-advanced acceleration of technological progress (some attributable to AI) that catalyzes a greater-than-human intelligence. Based on the success of human-brain interfaces, Kurzweil predicts that, by the year 2045, an AI-triggered transformation will forever alter the global economy and human civilization itself. In a follow-up published this year, *The Singularity Is Nearer*, Kurzweil reiterated this expectation as well as his belief that AI applications will realize human-level intelligence, which is the ability to perform on par with “the most skilled humans in a particular domain,” by 2029 in most respects.<sup>68</sup>

After taking a step back from the metaphysical, it is important to recognize that technological innovation is indeed accelerating, but that constraints remain. For example, today's AI excels at tasks that require sophisticated pattern recognition operations. These systems are almost certain to continue to improve.

However, our ability to develop AI systems that blend creativity, analysis, and judgment—such as systems with the ability to pair divergent and convergent thinking—is uncertain.

Given these realities, what will progress toward AGI realistically look like? Over the past decade, AI research has achieved breakthroughs largely by focusing on solving narrowly defined problems in areas such as object detection, video analysis, search and summarization, data mining, intelligent robotics, time series, and text mining. Future jumps in AI will continue through cumulative gains in solving other narrow problems.

These challenges could include causal or counterfactual reasoning (e.g., hypothesizing about the outcome to a possible event not yet executed); automation of work in a highly dynamic, multisensory environment; and scaling of AI hardware by separating storage and compute, akin to the use of a memristor in neuromorphic computing.

Such breakthroughs—especially when joined with faster deployment, testing, and updating of AI algorithms due to improvements in hardware and access to truly voluminous data—will likely lead, one future day, to a technology that might achieve comparable human performance at certain tasks. At the same time, we cannot lose sight of the fact that AGI's risks are real. They include almost certain workforce disruption, further threats to societal trust, and the potential to shift the balance of power dramatically and create technology that directly threatens humanity.

As AI continues to evolve, it will be critical to remember that this technology was built by humans, and our insight powers it. Therefore, we must continue to expect and ensure transparency and accountability in its development and operations. At some point soon, we are even likely to demand a new Turing test for computers, which is their ability to describe, demonstrate, and defend their reasoning.

“At some point, we really will have AGI. Maybe OpenAI will build it. Maybe some other company will build it.”

**Ilya Sutskever**, Chief Scientist of OpenAI, 2023<sup>69</sup>

# Acknowledgments

This report builds on the collective insight and wisdom of Booz Allen's AI practice with the following individuals playing a leadership role in its development:

**Lead Author:** Catherine Ordun, Ph.D.

**Coauthors:** Alexa Hoffman, Edward Raff, Ph.D., and Alison Smith

**Contributors:** Ryan Ashcraft, Bryan Castle, Drew Farris, Kate Helfet, Matt Keating, Chu Lahlou, Justin Neroda, and Ryan Swope

**Lead Editor:** Jim McDermott, Ph.D.

**Editorial Lead:** John Conley

We also recognize the authors of the original Artificial Intelligence Primer as key inspirations for this effort. They include J.D. Dulny, Ph.D., Josh Elliot, Drew Farris, Emma Kinnucan, Steve Mills, and Joshua Sullivan.



# Glossary of Terms

(Developed by ChatGPT)

**Agent:** An autonomous software entity designed to perform tasks or make decisions based on input, often using advanced algorithms like natural language processing and deep learning to interact with users and generate content.

**Alignment:** Ensuring that an AI system's goals, behaviors, and values are consistent with human values and intended outcomes, preventing it from acting in ways that are harmful or unintended.

**Backpropagation:** A training algorithm for neural networks that involves adjusting the weights of the connections based on the error rate of the output, allowing the network to learn by minimizing the difference between predicted and actual results.

**Convolutional Neural Network (CNN):** A type of deep learning algorithm primarily used for processing and analyzing visual data, such as images and videos, by mimicking the way human visual processing works.

**Diffusion Model:** A generative model that simulates the diffusion process to transform random noise into structured data. It is often used for tasks like image generation.

**Explainable AI:** An area of AI research focused on making machine learning models more interpretable and understandable. The goal is to provide transparency on how decisions are made.

**Few-Shot Learning in LLM:** Using an LLM to perform a task by providing it with a very small number of examples. It bridges the gap between zero-shot and extensive training.

**Fine-Tuning:** Adjusting the parameters of a pre-trained model on a smaller, task-specific dataset. This helps adapt the model to specific applications.

**Foundational Model:** A large, general-purpose pre-trained model that can be fine-tuned for specific tasks. It serves as a starting point for many AI applications.

**Generalizability:** The ability of a model to perform well on unseen data. It indicates how well a model can adapt to new situations.

**Generative Adversarial Network (GAN):** A type of neural network model comprising two networks (generator and discriminator) competing against each other. The generator creates data while the discriminator evaluates its authenticity.

**Generative Pre-Trained Transformer (GPT) Model:** A type of LLM developed by OpenAI. It is designed to generate human-like text based on given prompts.

**Large Language Model (LLM):** A type of neural network trained on vast amounts of text data. Examples include GPT models.

**Loss (Cost) Function:** A mathematical formula that measures the difference between the predicted output and actual data. It is used during training to adjust a model's weights.

**LLM “Chaining”:** Linking multiple prompts and responses together in a sequence to guide an LLM through more complex tasks or reasoning.

**Model Capacity:** Refers to the amount of information or complexity a model can capture. Higher capacity can mean better fit to data but also potential overfitting.

**Multimodality:** The ability of a model to process and integrate information from multiple types of data, such as text, images, audio, and video, to enhance understanding and performance across different tasks.

**Neural Network:** A computational model inspired by biological neurons used for tasks like classification, regression, and generation.

**Noise:** Random or unwanted fluctuations in data. In generative models, noise can serve as a starting point for generating data.

**Orchestration:** The coordinated management and automation of multiple AI models, processes, and workflows to achieve a seamless and efficient overall operation.

**Parameters or Weights:** Values in a neural network that are adjusted during training. They determine how input data is transformed into outputs.

**Probability Distribution:** A mathematical function that describes the likelihood of different outcomes. In AI, it is often used to model uncertainties.

**Prompt:** A question or statement given to an AI model to generate a response. It acts as the initial input to guide the model's output.

**Prompt Engineering:** The practice of crafting and refining prompts to optimize an AI model's response. This can help improve the quality or specificity of model outputs.

**Recurrent Neural Network (RNN):** A type of neural network designed for processing sequential data by maintaining a “memory” of previous inputs, making it suitable for tasks like language modeling and time series analysis.

**Reinforcement Learning from Human Feedback (RLHF):** A method where an AI system is trained to make decisions using feedback from human evaluators, enhancing its ability to align with human values and preferences.

**Reproducibility:** The ability to consistently reproduce the same results using the same data and methods. It ensures the reliability of experiments and studies.

**Self-Attention:** A mechanism in neural networks that weighs input data differently, enabling the model to focus on more relevant parts. It's a key component of Transformer architectures.

**Supervised Machine Learning:** A learning approach where a model is trained on labeled data, meaning each example in the training dataset is paired with the correct output.

**Token:** A unit of text resulting from tokenization. It can represent a word, a part of a word, or even a character.

**Tokenization:** The process of converting a sequence of text into smaller units (tokens) for easier processing. Common tokens include words or sub-words.

**Transfer Learning:** Using a model trained on one task as the foundation for training on a different, but related, task. This approach leverages previously learned knowledge.

**Transformer Model:** A type of neural network architecture known for self-attention mechanisms. Widely used in tasks like machine translation and NLP.

**Unsupervised Machine Learning:** A learning approach where a model is trained without labeled data. The goal is often to discover patterns or structures within the data.

**Zero-Shot Learning in LLM:** Using an LLM to perform a task without any specific examples or training on that task. The model leverages general knowledge.

# Endnotes

- 1 “The AI Dilemma,” AI for Humanity, O’Reilly Media. <https://www.oreilly.com/library/view/ai-for-humanity/9781394180301/c04.xhtml>.
- 2 “Booz Allen Ranks First in Artificial Intelligence Services.” <https://www.boozallen.com/insights/ai/booz-allen-ranks-first-in-artificial-intelligence-services.html>.
- 3 Bengio, Yoshua. “Deep Learning of Representations for Unsupervised and Transfer Learning.” Proceedings of ICML Workshop on Unsupervised and Transfer Learning. 2012.
- 4 Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv preprint. arXiv:1409.0473. 2014.
- 5 Vaswani, Ashish, et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems 30. 2017.
- 6 Schulman, John, et al. “Proximal Policy Optimization Algorithms.” arXiv preprint. arXiv:1707.06347. 2017.
- 7 Turing, Alan M. “Computing Machinery and Intelligence.” Mind 59(236): 433–460.
- 8 Fahlman, Scott E., and Geoffrey E. Hinton. “Connectionist Architectures for Artificial Intelligence.” Computer 20.01. 1987. 100–109.
- 9 Rumelhart, David E., et al. “Backpropagation: The Basic Theory.” Backpropagation. Psychology Press, 2013: 1–34.
- 10 LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet. “Convolutional Networks and Applications in Vision.” Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE. 2010.
- 11 LeCun, Yann, et al. “Backpropagation Applied to Handwritten Zip Code Recognition.” Neural Computation 1.4. 1989: 541–551.
- 12 Mikolov, T., Chen, K., Corrado, G., & Dean, J. “Efficient Estimation of Word Representations in Vector Space.” In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, 2013.
- 13 Christopher Olah. “Understanding LSTM Networks.” Colah’s Blog. August 27, 2015.
- 14 Hochreiter, Sepp, and Jürgen Schmidhuber. “Long Short-Term Memory.” Neural Computation 9.8. 1997: 1735–1780.
- 15 Mikolov, Tomas, et al. “Distributed Representations of Words and Phrases and Their Compositionality.” Advances in Neural Information Processing Systems 26. 2013.
- 16 Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III, 18. Springer International Publishing, 2015.
- 17 Van Veen, Fjodor. “The Neural Network Zoo.” The Asimov Institute Blog. September 14, 2016.
- 18 François Chollet. <https://twitter.com/fchollet/status/951906139632840704?lang=en>.
- 19 <https://quebecartificialintelligence.com/academy/>
- 20 Janocha, Katarzyna, and Wojciech Marian Czarnecki. “On Loss Functions for Deep Neural Networks in Classification.” arXiv Preprint. arXiv:1702.05659 (2017).
- 21 Shannon, C.E. “A Mathematical Theory of Communication.” The Bell System Technical Journal. 1948. <https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>.
- 22 A Combination of Softmax Regression that Generalizes Logistic Regression into a Multi-Class Problem, with Negative Log Likelihood.
- 23 Rothman, Joshua. “Why the Godfather of AI Fears What He’s Built.” The New Yorker, November 2023.
- 24 Rumelhart, David E., et al. “Backpropagation: The Basic Theory.” Backpropagation. Psychology Press, 2013: 1–34.
- 25 Li, Katherine (Yi). “Vanishing and Exploding Gradients in Neural Network Models: Debugging, Monitoring, and Fixing.” Neptune.AI Blog. August 16, 2024.
- 26 Ordun, Catherine. “A Brief Overview of GANs” in Multimodal Deep Generative Models for Cross-Spectral Image Analysis. University of Maryland, Baltimore County, 2023.
- 27 Silver, David. “Deep Reinforcement Learning.” <https://deepmind.google/discover/blog/deep-reinforcement-learning/>.
- 28 Metz, David. “In Two Moves, AlphaGo and Lee Sedol Redefined the Future.” Wired. 2016. <https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/>.
- 29 Tosato, Giulio et al. “EEG Synthetic Data Generation Using Probabilistic Diffusion Models.” arXiv Preprint. arXiv:2303.06068. March 2023.
- 30 Rombach, Robin, et al. “High-Resolution Image Synthesis with Latent Diffusion Models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
- 31 Chen, Nanxin, et al. “Wavegrad: Estimating Gradients for Waveform Generation.” arXiv Preprint. arXiv:2009.00713 (2020).
- 32 Shang, Yuzhang, et al. “Post-Training Quantization on Diffusion Models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- 33 Cottier, Ben, et al. “How Much Does It Cost to Train Frontier AI Models?” Epoch AI. 2024. <https://epochai.org/blog/how-much-does-it-cost-to-train-frontier-ai-models>.
- 34 Ouyang, Long, et al. “Training Language Models to Follow Instructions with Human Feedback.” Advances in Neural Information Processing Systems 35. 2022: 27730–27744.
- 35 Leike, Jan, et al. “Scalable Agent Alignment via Reward Modeling: a Research Direction.” arXiv Preprint. arXiv:1811.07871. 2018.
- 36 Christiano, Paul F., et al. “Deep Reinforcement Learning from Human Preferences.” Advances in Neural Information Processing Systems 30. 2017.
- 37 Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv Preprint. arXiv:1409.0473. 2014.



- 38 Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv Preprint. arXiv:1409.0473. 2014.
- 39 Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. 2017.
- 40 Vaswani, Ashish et al. "Attention is All You Need." 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- 41 Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2. 2018: 423-443.
- 42 Ordun, Catherine, et al. "Intelligent Sight and Sound: a Chronic Cancer Pain Dataset." *NeurIPS 2022 Datasets and Benchmarks Track*.
- 43 Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning*. PMLR, 2021.
- 44 Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv Preprint. arXiv:2010.11929. 2020.
- 45 Hu, Edward J., et al. "Lora: Low-Rank Adaptation of Large Language Models." arXiv Preprint. arXiv:2106.09685. 2021.
- 46 Weng, Lilian. "LLM Powered Autonomous Agents." <https://lilianweng.github.io/posts/2023-06-23-agent/>.
- 47 Weng, Lilian. "LLM Powered Autonomous Agents." <https://lilianweng.github.io/posts/2023-06-23-agent/>.
- 48 Nilsson, Nils J. "Principles of Artificial Intelligence, First Edition." Morgan Kaufman. February 15, 1982.
- 49 Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore. "Reinforcement Learning: A Survey." *Journal of Artificial Intelligence Research* 4. 1996: 237-285.
- 50 Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., and Lowe, R. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
- 51 Raschka, Sebastian. "LLM Training: RLHF and Its Alternatives." <https://magazine.sebastianraschka.com/p/llm-training-rlhf-and-its-alternatives>; Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., and Lowe, R. (2022). "Training Language Models to Follow Instructions with Human Feedback,". *Advances in Neural Information Processing Systems*, 35, 27730-27744.; Lambert, et al., "Illustrating Reinforcement Learning from Human Feedback (RLHF)," *Hugging Face Blog*. 2022.
- 52 Schulman, John, et al. "Proximal Policy Optimization Algorithms." arXiv Preprint. arXiv:1707.06347. 2017.
- 53 Schulman, John, et al. "Proximal Policy Optimization Algorithms." arXiv Preprint. arXiv:1707.06347. 2017.
- 54 Narayanan, Deepak, et al. "Scaling Language Model Training." <https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/>.
- 55 Jiang, Biao, et al. "MotionGPT: Human Motion as a Foreign Language." arXiv Preprint. arXiv:2306.14795. 2023.
- 56 Ordun, Catherine, et al. "A Generative Approach for Image Registration of Visible-Thermal (VT) Cancer Faces." *MICCAI Workshop on Artificial Intelligence over Infrared Images for Medical Applications*. Cham: Springer Nature Switzerland, 2023.
- 57 Jiang, Biao, et al. "Motiongpt: Human motion as a foreign language." *Advances in Neural Information Processing Systems* 36 (2023): 20067-20079.
- 58 Ordun, Catherine, et al. "A Generative Approach for Image Registration of Visible-Thermal (VT) Cancer Faces." *MICCAI Workshop on Artificial Intelligence over Infrared Images for Medical Applications*. Cham: Springer Nature Switzerland, 2023.
- 59 Ordun, Catherine, Edward Raff, and Sanjay Purushotham. "Vista Morph-Unsupervised Image Registration of Visible-Thermal Facial Pairs." *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023.
- 60 Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35.
- 61 Yadav, Abhay, et al. "Stabilizing Adversarial Nets with Prediction Methods." arXiv Preprint. arXiv:1705.07364. 2017.
- 62 Nguyen, Tu, et al. "Dual Discriminator Generative Adversarial Nets." *Advances in Neural Information Processing Systems* 30. 2017.
- 63 Metz, Luke, et al. "Unrolled Generative Adversarial Networks." arXiv Preprint. arXiv:1611.02163. 2016.
- 64 Özbey, Muzafer, et al. "Unsupervised Medical Image Translation with Adversarial Diffusion Models." *IEEE Transactions on Medical Imaging* (2023); Ordun, Catherine, Edward Raff, and Sanjay Purushotham. "When Visible-to-Thermal Facial GAN Beats Conditional Diffusion." arXiv Preprint. arXiv:2302.09395. 2023.
- 65 Shumailov, Ilia, et al. "The Curse of Recursion." arXiv Preprint. arXiv:2305.17493.
- 66 Shumailov, Ilia et al. "AI Models Collapse When Trained on Recursively Generated Data." *Nature*. 631 (8022): 755–759. doi:10.1038/s41586-024-07566-y. ISSN 1476-4687. PMC 11269175. July 2024.
- 67 Heaven, Will. "Rogue Superintelligence and Merging with Machines." *MIT Technology Review*. 2023. <https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/>.
- 68 Goertzel, Ben. "Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's Critique of Kurzweil." *Artificial Intelligence* 171.18. 2007. 1161-1173.
- 69 <https://web.archive.org/web/20180410074243/http://mindstalk.net/vinge/vinge-sing.html>.
- 70 Kurzweil, Ray. "The Singularity Is Near." *Ethics and Emerging Technologies*. London: Palgrave Macmillan UK, 2005. 393-406.
- 71 Corbyn, Zoe. "AI Scientist Ray Kurzweil." *The Guardian*. <https://www.theguardian.com/technology/article/2024/jun/29/ray-kurzweil-google-ai-the-singularity-is-nearer>.
- 72 Yang, Jingfeng, et al. "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond." arXiv Preprint. arXiv:2304.13712. 2023.



## About Booz Allen

Booz Allen is the advanced technology company delivering outcomes with speed for America's most critical defense, civil, and national security priorities. We build technology solutions using AI, cyber, and other cutting-edge technologies to advance and protect the nation and its citizens. By focusing on outcomes, we enable our people, clients, and their missions to succeed—accelerating the nation to realize our purpose: Empower People to Change the World®.

**[BoozAllen.com/AIPrimer](https://BoozAllen.com/AIPrimer)**