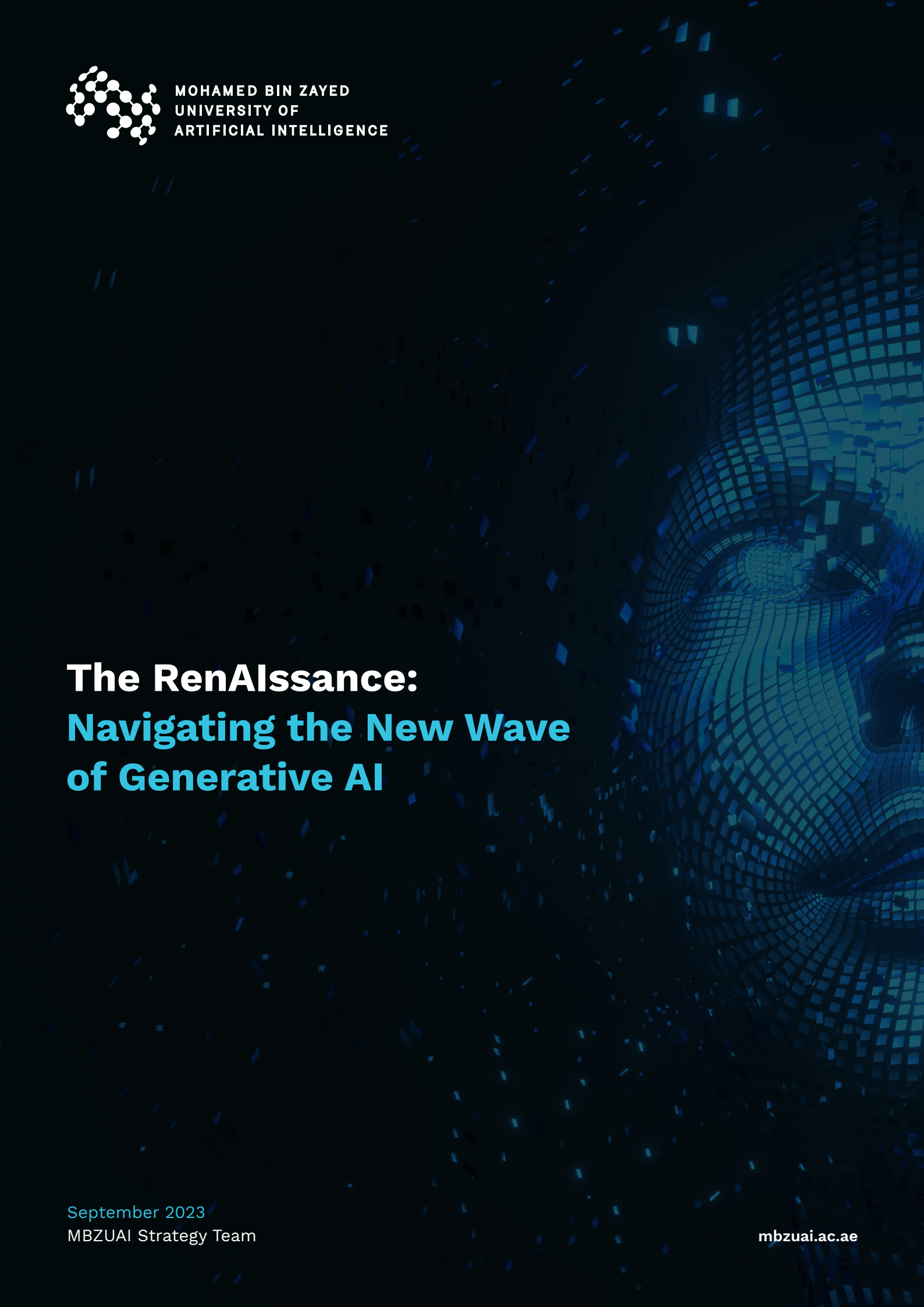
MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE



# The RenAIssance:
# Navigating the New Wave of Generative AI

September 2023
MBZUAI Strategy Team

mbzuai.ac.ae

# Executive summary

## Section 1
## Navigating the generative AI revolution: a new dawn of technological possibilities

The advent of ChatGPT, a groundbreaking application, marks our entry into the realm of generative AI capable of comprehending human language and crafting unique content. This capability is driven by foundation models, which are designed to assimilate information from diverse modalities and be tailored for numerous applications. These models emerge from processing vast datasets through a transformer model, enabling them to generate content, understand context, and reason logically.

Tech giants such as Microsoft, Meta, and Google have been at the forefront of pioneering foundation models. Simultaneously, other notable tech entities and emerging startups, including Anthropic, Inflection AI, and Cohere, have significantly advanced their development.

The introduction of foundation models has revolutionized AI capabilities, enabling a plethora of applications across diverse sectors such as marketing, customer support, productivity tools, and risk management. The versatility of these models allows for their integration into myriad settings, underscoring their transformative potential. In fact, adoption of generative AI has already been widespread across multiple industries. It's projected that foundation models could automate nearly 25% of all work tasks, enhancing productivity and reducing labor costs.

The ascendancy of generative AI has reshaped the technological landscape, delineating distinct stages in its value chain and unlocking numerous market opportunities. This transformative wave has spurred a surge in startups specializing in the domain, with investment in the first half of 2023 surging to five times that of the entire previous year.

## Section 2
## Essentials for building foundation models

The cornerstones of creating a foundation model are data, computational capacity, and expertise.

**Data:** The performance of foundation models hinges on the scalability, quality, and diversity of data that they are trained on. Most models utilize diverse public textual datasets for pre-training.

**Compute:** Crafting large-scale models necessitates substantial hardware and computational resources. For instance, training of OpenAI's GPT-3 took 34 days on 1,024 graphics processing unit (GPUs), with an estimated cost of $4.6 million in compute alone. However, evolving hardware technologies promise to bring down these costs.

**Talent:** Assembling a foundation model necessitates a robust team of 70-80 experts. The involvement of top-tier industry experts with extensive experience in large-scale machine learning projects is crucial.

## Section 3
## Future horizons of generative AI: scaling, alignment, and multimodality

**We anticipate three primary trajectories for generative AI:**

**Scaling up:** The industry is on an audacious journey towards creating models that are grander in both scale and capability. Yet, due to hardware constraints, the next generation of models might only be two or three times of the size of their current counterparts.

**AI Alignment and evaluation:** The alignment of AI models with human values, ethics, and objectives remains a top priority. At the same time, rigorous evaluations of their performance, fairness, and societal implications are essential to cultivate trust in AI. Presently, the current AI alignment practice heavily relies on reinforcement learning from human feedback (RLHF). Additionally, Anthropic has introduced the "constitutional AI" approach, aiming for scalable oversight. Yet, achieving true AI alignment demands a holistic approach and requires collaboration across diverse disciplines and industries.

**Multi-modality:** Recently, a key development in the AI realm is the fusion of foundation models with robotics. This convergence is drawing attention from both industry titans and budding startups, heralding a future where AI seamlessly interacts with the physical world. Tech powerhouses such as Microsoft, Meta, and Tesla are venturing into this space.

# Introduction

**The landscape of artificial intelligence (AI) has undergone a remarkable transformation in recent years, captivating the imagination of the world. It was in 2016 when AlphaGo, an AI program crafted by DeepMind, astounded the world by defeating a reigning world champion Go player, marking a significant turning point in the public's perception of AI. Fast forward just six years, and we find ourselves in a new era of technological wonder. The launch of GPT-4 by OpenAI demonstrated the boundless possibilities of artificial intelligence and marks the beginning of an era of generative AI and foundational models. These models, with their unparalleled ability to understand, generate, and reason with human language, are not just technological marvels but harbingers of a future where AI seamlessly integrates into every facet of our lives.**

In pursuit of future readiness and in alignment with the visionary directives of its leadership, the United Arab Emirates (UAE) has been at the forefront of embracing innovation and emerging technologies. The UAE National Strategy for Artificial Intelligence 2031 has set a clear objective to position the nation as the global leader in AI by the year 2031. To achieve this ambitious goal, it is imperative for us all to embark on a journey of understanding the power, reach, and capabilities of generative AI.

This report serves as a gateway for the reader to gain a fundamental understanding of generative AI and its far-reaching impacts. Emphasizing the business aspects of generative AI, it delves into the nuances of model development, highlights the generative AI value chain, and demonstrates its potential to boost business efficiency and improve quality of life. The report concludes with a forward-looking perspective on how this revolutionary technology will evolve and persistently transform our world.

01

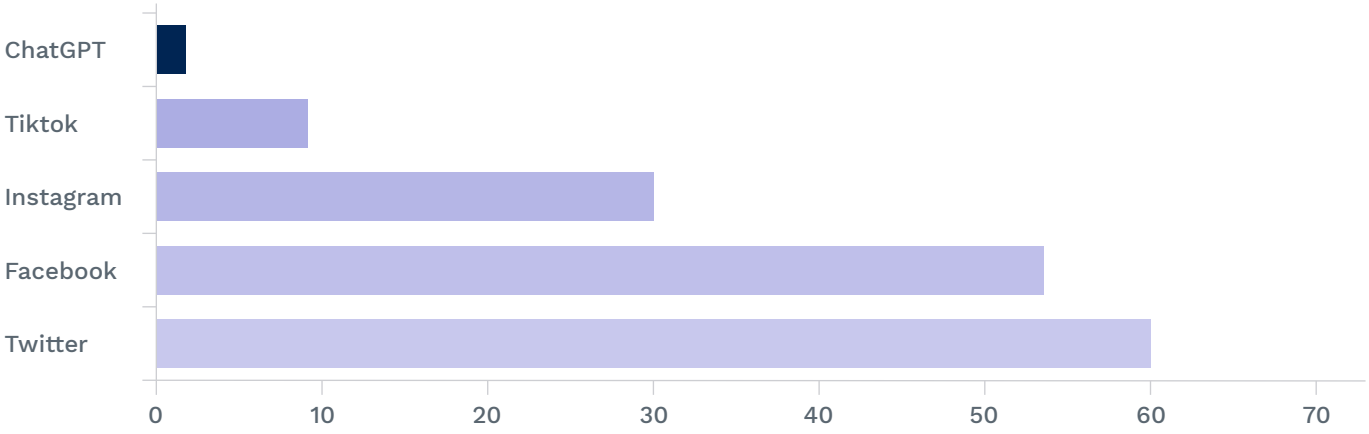
# Navigating the generative AI revolution:

**The digital epoch is being redefined by the meteoric rise of generative AI. Pioneered by platforms such as ChatGPT, these technologies are not merely tools but catalysts, reshaping the contours of sectors ranging from finance to healthcare. As generative AI seamlessly embeds itself in our daily routines and business operations, it is poised not only to complement human endeavors, but also to set new benchmarks in innovation, efficiency, and potential.**

## 1.1 Generative AI: the new frontier

The emergence of ChatGPT, developed by OpenAI, represents a pivotal moment in the field of generative AI. This technology, which can both understand human language and produce high-quality original content, has seen a huge rise in its adoption. Launched for public testing in November 2022, ChatGPT's user base expanded rapidly, reaching 100 million monthly active users (MAU) within a mere two months. This impressive growth outpaced even some of the most disruptive platforms in recent memory – including TikTok, Instagram, and Facebook – positioning ChatGPT as one of the most rapidly adopted applications in history.

**Time to reach 100 million MAU (months)**




ChatGPT reached 100 million MAU in two months, becoming the fastest growing application in history
**Source:** MBZUAI research

At its core, generative AI uses advanced algorithms to create diverse forms of content, such as text, images, or other multimedia. These algorithms are powered by foundation models – large AI systems trained on a vast quantity of input data at scale, enabling them to process and generate content that lies beyond the training data. The versatility of these models is remarkable; they can handle a myriad of data types, from text and images to videos.

For instance, ChatGPT, which primarily focuses on text, is built upon a large language model (LLM), allowing it to both interpret and produce textual content. Midjourney is a text-to-image AI that employs both a large language model and a diffusion model (which generates high-quality images from textual descriptions) to translate natural language prompts into visuals. Once these foundational models are trained, they can be fine-tuned for a broad spectrum of applications. This includes tasks such as conversation, sentiment analysis, image captioning, object recognition, among others.

**Several leading technology companies have pioneered advancements in foundation model development, each showcasing distinct innovations and areas of expertise. These innovations range from intricate image generation from textual descriptions and advanced conversational interfaces to groundbreaking techniques in bioinformatics prediction.**

---

**OpenAI,** with backing from Microsoft, introduced the world to ChatGPT in 2022 and promptly unveiled the more sophisticated GPT-4 in early 2023. Another significant contribution from OpenAI was the launch of DALL·E 2 in April 2022. This model, an evolution of the original DALL·E from 2021, possesses the remarkable ability to generate intricate images based on textual descriptions. Both models are closed source, with their code unavailable to the public. However, they can be accessed via APIs, and usage is primarily governed by licensing agreements.

---

**DeepMind,** operating under the Google umbrella, garnered widespread attention in 2021 with the release of AlphaFold. This open-source tool predicts protein structures based purely on amino acid sequences. By May 2023, DeepMind further expanded its generative AI portfolio by introducing several new features, one of which is the new LLM named PaLM 2, designed to enhance their Bard chatbot. PaLM 2 is closed source, with access limited to research purposes only.

---


**Meta** marked its presence in the generative AI space by launching the LLaMA language model in February 2023, followed by the release of LLaMA 2 in July. Unlike its predecessor, LLaMA 2 is an open-source project, free for research and commercial use. This move democratizes access, enabling a wider community to use, distribute, and even modify their model codes.

---

**The generative AI landscape is not solely dominated by these technology giants. Emerging startups such as Cohere and Anthropic are also making significant advancements. These newer players are pushing the boundaries of generative AI capabilities, notably promoting ethical and responsible deployment of such technologies.**

---

**Anthropic,** founded by former members of OpenAI, aims to build frontier AI systems that are reliable, interpretable, and steerable. They launched Claude 2 in July 2023, differentiating itself from competing models such as ChatGPT by focusing on enterprise and security. On AI alignment, Anthropic employs reinforcement learning from AI feedback – instead of reinforcement learning from human feedback – to improve safety and reduce harm.

---

**Inflection AI**, a machine learning start-up founded in 2022, aims to create a "personal AI for everyone". Inflection's flagship product, Pi, is an AI-driven assistant tailored to deliver information aligned with an individual's interests and needs. Accessible through messaging apps or online, Pi is designed to be a compassionate companion, providing advice and insights in a congenial and fluid manner. In June 2023, Inflection secured a staggering $1.3 billion in funding, from tech giants such as Microsoft and Nvidia, as well as industry stalwarts such as Reid Hoffman, Bill Gates, and Eric Schmidt.

---

**Cohere** is dedicated to crafting machines that comprehend the world, making them safely accessible to everyone. With a vision to revolutionize enterprises through AI, Cohere's advanced models facilitate interactive chat functionalities, generate descriptive text for products, blogs, and articles, and discern the essence of text for various applications such as search and intent recognition. In July 2023, Cohere entered a strategic alliance with McKinsey, aiming to leverage generative AI to offer bespoke solutions that enhance business performance. Furthering their innovations, Cohere unveiled "Coral" in August, a knowledge assistant designed to bolster enterprise productivity.

---

## 1.2 The multifaceted applications

**Foundation models have ushered in a new era of AI capabilities, distinguishing themselves from previous AI technologies. Their enhanced features provide a robust foundation, making them powerful assets in diverse applications.**

**Versatility:**

Beyond traditional tasks, these models are adept at a spectrum of functions, from answering intricate questions to generating original content.

**Contextual understanding:**

Foundation models stand out in their ability to analyze inputs in relation to their context, ensuring that the responses generated are both accurate and contextually relevant.

**Adaptive Learning:**

As they interact and receive feedback, foundation models refine their processes, thereby improving their accuracy and dependability over time.

**Generalization:**

Drawing from their extensive training data, these models can extrapolate and apply their knowledge to handle new and distinct tasks effectively.

**Owing to their distinctive capabilities, generative AI models have found applications across a wide spectrum of domains, underscoring their transformative potential:**

### Marketing and sales:

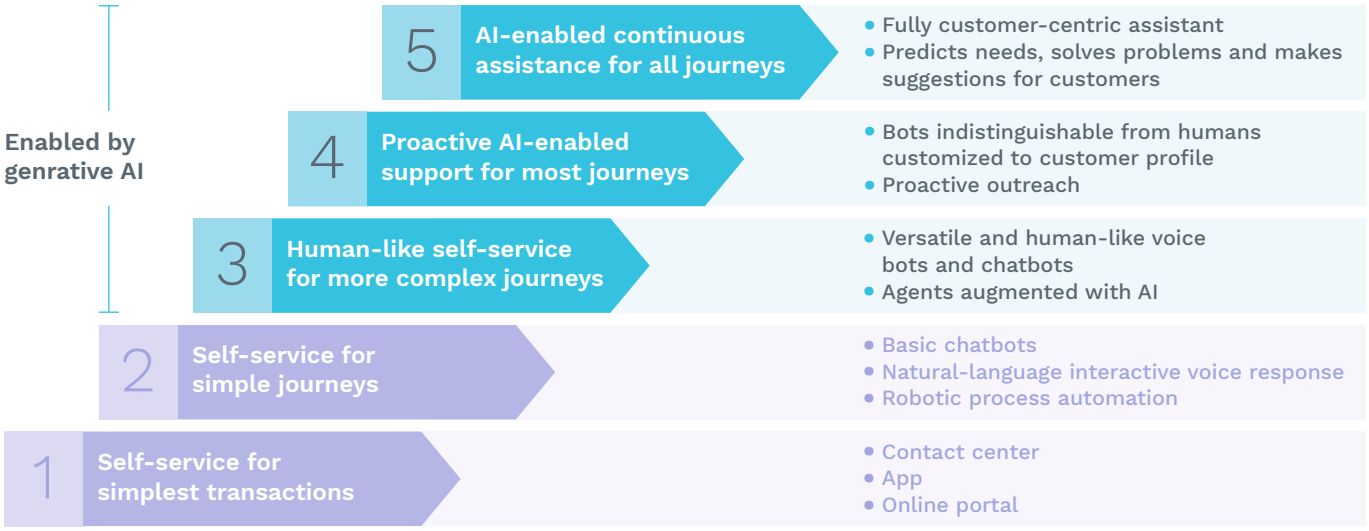
Generative AI has the ability to automate content creation, enhance existing content, and generate visuals. Beyond generic content, it can create personalized messages based on individual customer inclinations and actions.

*Example: Twain offers an AI-driven sales coach technology specifically tailored to boost the conversion rates of outreach messages. Sales executives can input their messages into Twain's specialized editor and promptly receive suggestions grounded in proven outreach strategies.*

### Customer support:

Generative AI models effectively address intricate customer queries by analyzing conversations for context and sentiment, delivering coherent and relevant responses. In addition, by leveraging customer data, these models offer personalized recommendations and solutions to elevate the customer experience. Productivity in the field will potentially be boosted by 30-50%.

*Example: IBM's Watson Assistant, a conversational AI platform, allows businesses to automate customer service responses. This platform enables the creation of AI-driven voice agents and chatbots, ensuring seamless self-service support across various communication channels*




AI-enabled customer service is maturing rapidly
**Source:** BCG

### Productivity tools:

Generative AI bolsters efficiency in various industries, automating administrative tasks, offering insights through data analysis, and enhancing content creation. It also proves beneficial in learning and development, simplifying complex concepts.

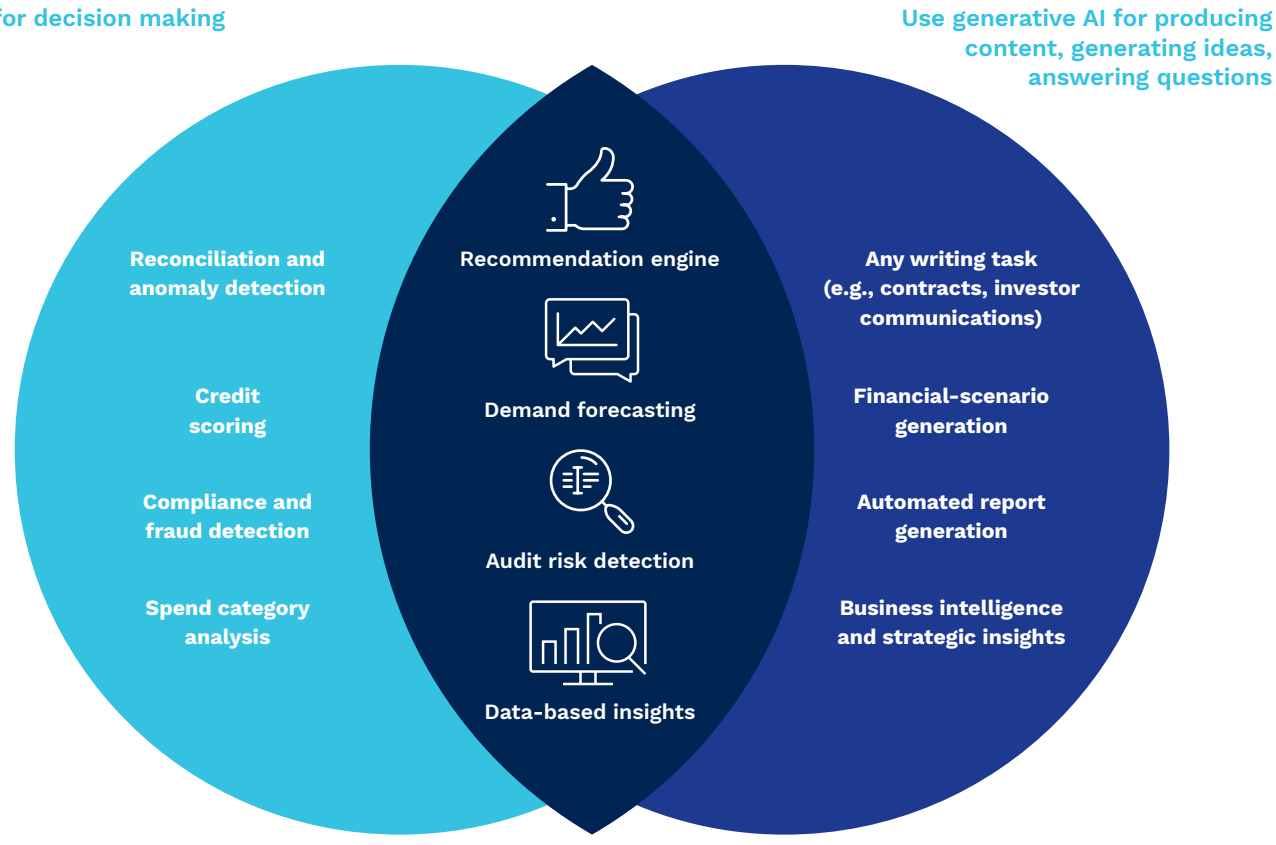
*Example: Copilot, introduced by Microsoft and powered by GPT-4, uses large language models to automate tasks within its portfolio of work apps including Word, Excel, PowerPoint, Outlook, and Teams. With Copilot, users can use generative AI to draft text in Word, create slideshows in PowerPoint from a simple prompt, and analyze data in Excel to generate charts and graphs.*

**Risk and legal assessment:**
Generative AI aids in automating legal document creation, contract reviews, and monitoring regulatory shifts, ensuring comprehensive risk assessments. For legal professionals, the technology also enhances lawyers' capabilities significantly by automating routine tasks and facilitating the creation of innovative legal theories.

*Example: By incorporating generative-AI-empowered features such as smart data ingestion and automated file indexing, Logikcull enables legal teams to navigate through extensive data with remarkable accuracy and swiftness. It also significantly minimizes the likelihood of missing vital legal data.*

**Strategy and finance:**
Generative AI is revolutionizing the finance sector as it streamlines operations and enhances accuracy in various functions. It enables the rapid development of forecasts and budgets, while also generating insightful financial commentary, and automating presentations. This technology is adept at harnessing public data for market intelligence, creating actionable customer and competitive insights. It can also extract strategic insights from CRM and ERP data, thereby influencing key financial and strategic decisions.

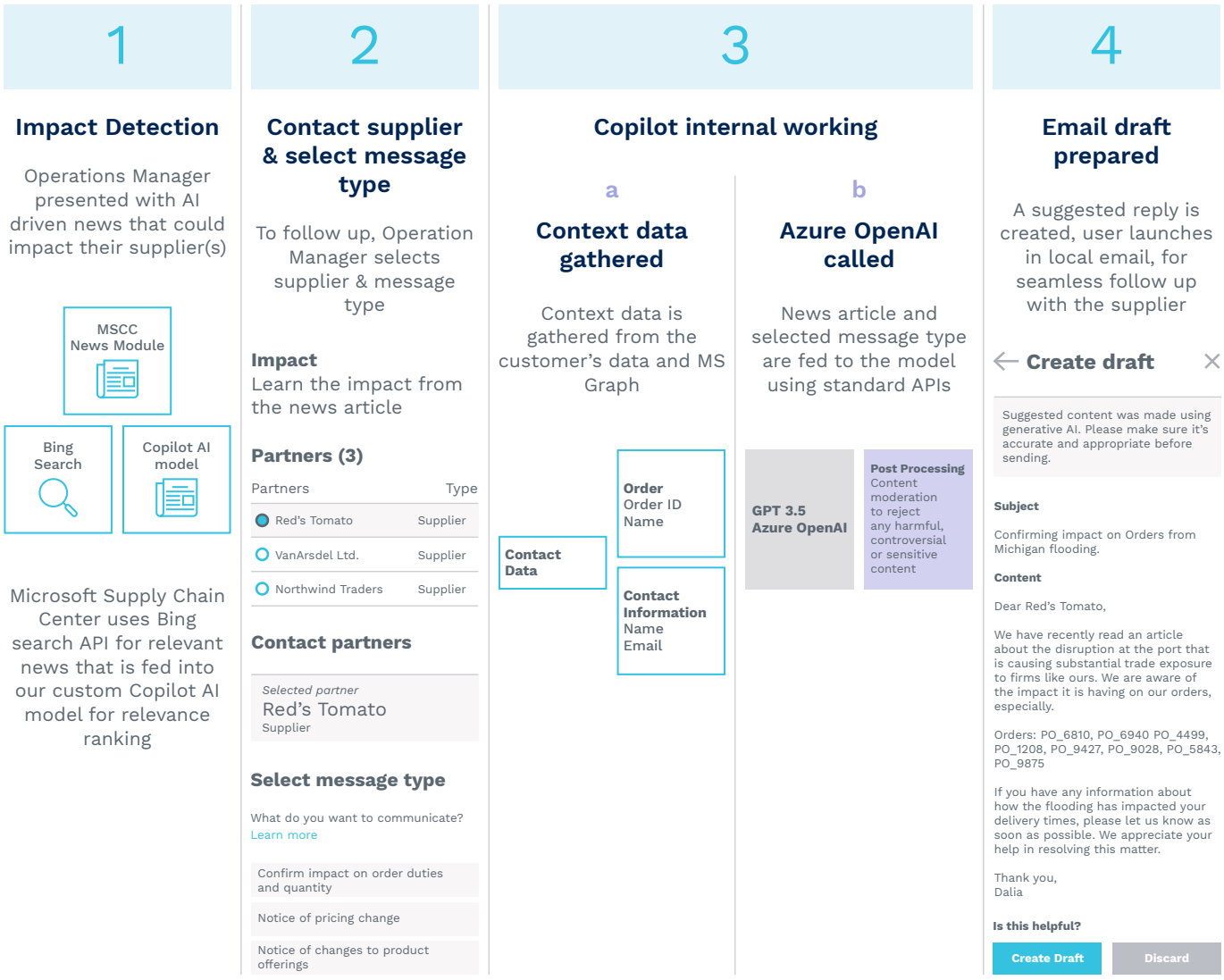
*Example: Morgan Stanley is set to introduce an advanced chatbot, powered by OpenAI's latest technology, to assist its financial advisors. This tool will empower advisors to query and analyze vast content and data volumes, with responses rooted in Morgan Stanley Wealth Management content, complete with source links.*




Generative AI and traditional AI have both separate and combined finance applications
**Source:** BCG

**Supply chain operations:**
AI can identify patterns and predict outcomes that can lead to more sustainable practices. It fosters data-driven decision-making, enabling swift adaptation to market changes and unveiling cost-saving strategies.

*Example: Microsoft utilizes generative AI in their Dynamics 365 to optimize demand forecasting through the analysis of historical sales data and external factors. This integration helps in honing inventory management and resource allocation, identifying efficient transportation routes, and mitigating supplier risks – thus saving costs and elevating operational efficiency.*




How supplier news communication is facilitated by copilot in Microsoft supply chain center
**Source:** Microsoft Dynamics 365

**Talent management:**
Generative AI can analyze data to draft job requirements and promote skills-based hiring. This technology can also elevate the employee experience by offering personalized solutions, fostering a growth-oriented work environment, and crafting individualized career paths.

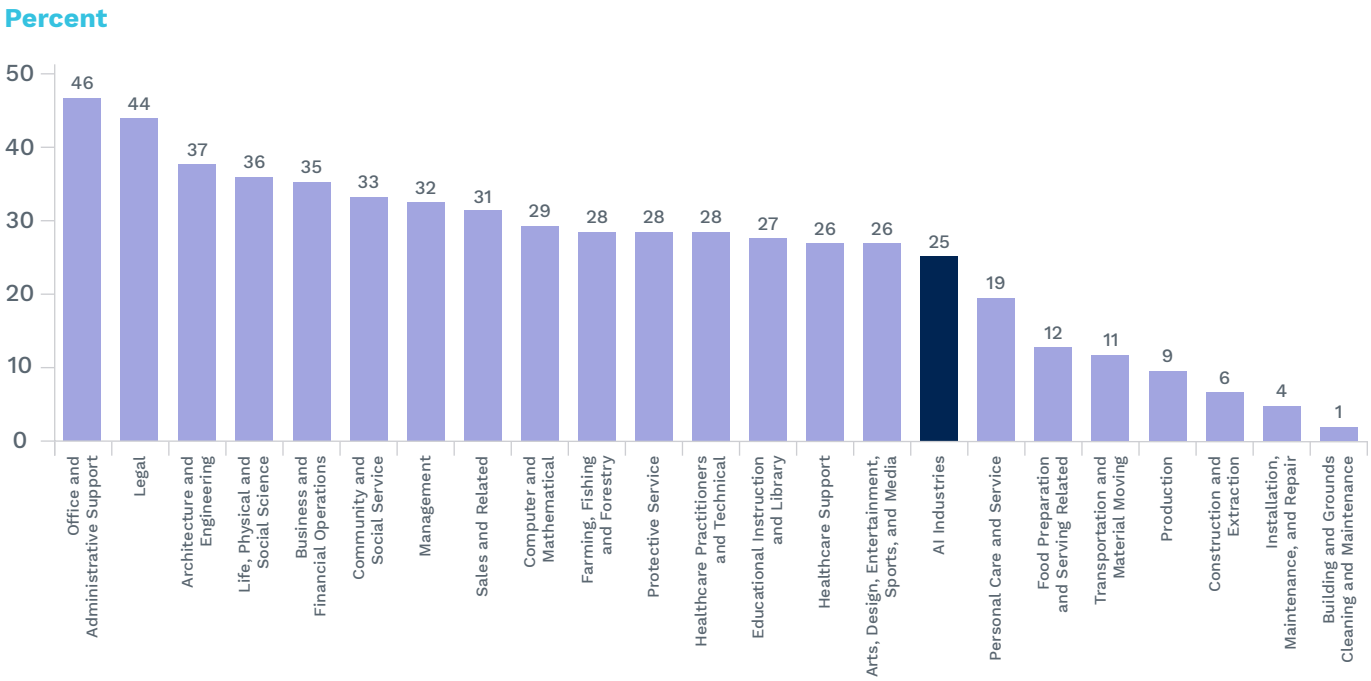
*Example: PeopleStrong's HR Tech platform integrates generative AI across the talent lifecycle, including performance, succession planning, and recruitment. A unique feature includes a talent coach that guides employees in career planning, skill identification, and offers curated courses for skill enhancement.*

**Engineering and product development:**
Generative AI facilitates rapid prototyping, creativity, code optimization, and task automation for developers, improving efficiency. This technology streamlines workflow, boosts productivity, and drives innovation while maintaining best practices.

*Example: GitHub Copilot, powered by a generative AI model developed by GitHub, OpenAI, and Microsoft, analyzes the context in the file that the users are editing and offers autocomplete-style suggestions.*

**Research and development:**
Generative AI helps R&D professionals spot potential innovation areas and predict trends by recognizing patterns and correlations. This ensures effective resource allocation and greater research productivity.

*Example: Biomap, specializing in innovative biotechnologies, leverages generative AI for enhanced research quality and analysis precision, enabling physicians to improve new drug efficacy. Their work includes developing generative proteins, predicting specific cell types and genes for disease modulation, and designing protein for precise therapy.*




A quarter of all work tasks are exposed to automation by generative AI
**Source:** Goldman Sachs Global Research Team

With their versatile application, foundation models are projected to achieve substantial boosts in productivity and significant reductions in labor costs for almost any professional activities. The most profound impacts of this technology are anticipated in domains such as office administration and support (46% percent exposed to automation), legal services (44% percent exposed to automation), and architecture and engineering (37% percent exposed to automation). Roles that inherently require a high degree of manual or human touchpoints might remain less influenced by this wave of automation (building, grounds cleaning and maintenance has only 1% exposure to automation).

Overall, it is estimated that one quarter of all work tasks are exposed to automation by Generative AI. One thing is clear: an increasing number of businesses are keenly integrating large language models into their operations, aiming to secure a competitive edge in a rapidly evolving marketplace.

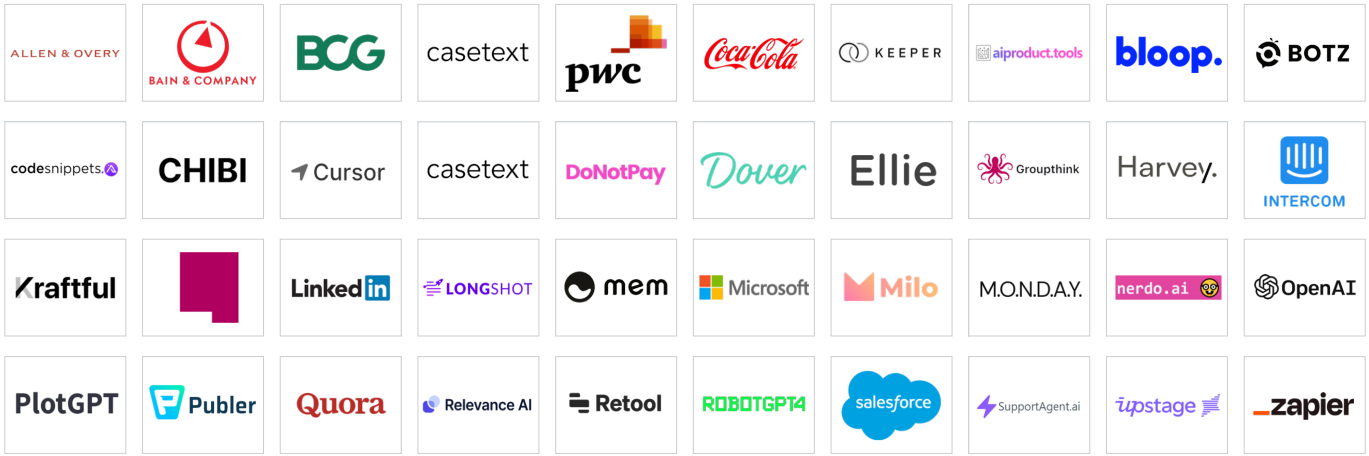
Today, the transformative potential of generative AI is already evident across multiple industries. For example, technology, finance, life sciences, and retail, are actively harnessing its capabilities.

- **Financial Services:** In early 2023, Bloomberg launched BloombergGPT, a large language model specifically trained on a wide range of financial data, to support a diverse set of natural language processing tasks within the financial industry. The research showed that BloombergGPT outperforms similarly-sized open models on financial NLP tasks by significant margins.

- **Life Science:** Moderna has been using machine learning and AI to advance mRNA-based vaccines and therapeutics across seven modalities. Now, in collaboration with IBM, they aim to harness generative AI for the optimal design of mRNA medicines, emphasizing safety and performance.

- **Retail:** eBay's ShopBot acts as a personal shopping guide, helping users sift through its billion-plus listings to find the best deals. Users can interact with ShopBot via text, voice, or photos, and the bot engages in further conversations to provide customized recommendations.

- **Software:** Adobe's Firefly, a suite of creative generative AI models, now boasts the Generative Fill feature, enabling users to augment images and manage objects with ease. Seamlessly integrated into Photoshop, Firefly combines the rapidity of generative AI with Photoshop's meticulous precision.




First 50 enterprise customers for OpenAI's GPT-4
**Source:** LifeArchitect.ai

## 1.3 The ecosystem

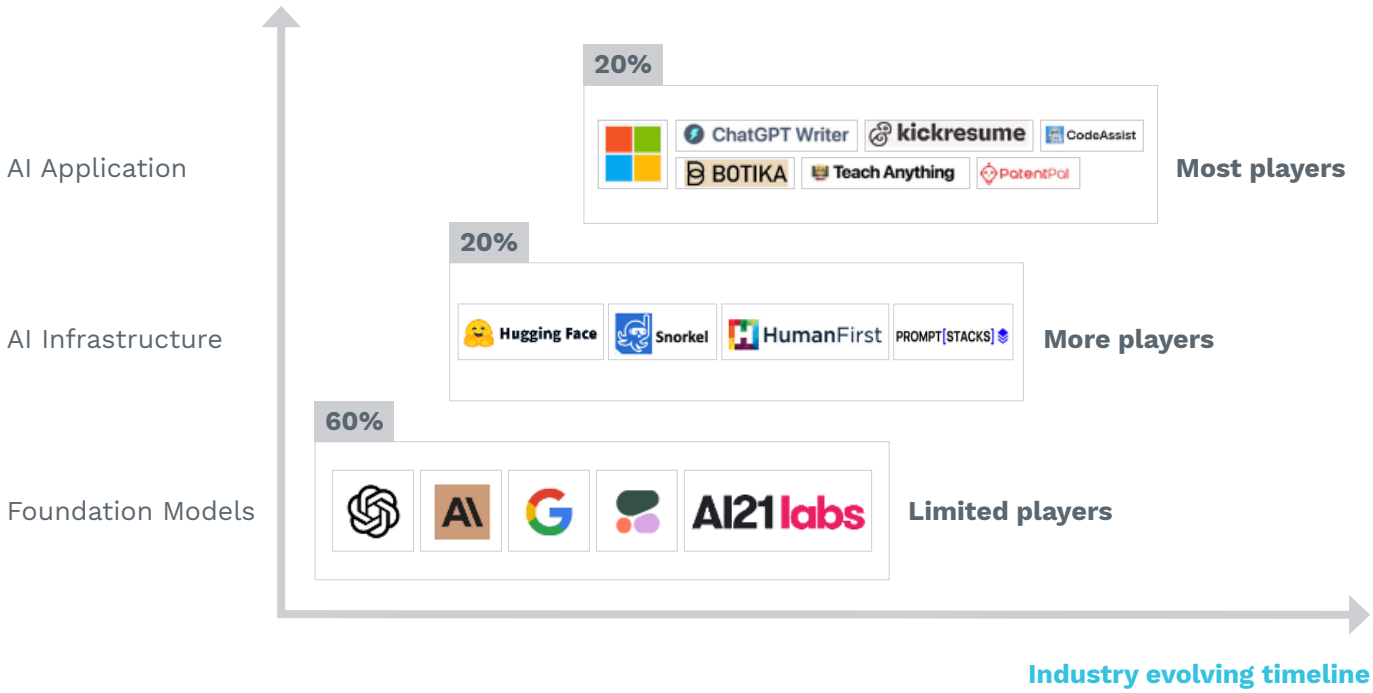
The growth of generative AI has changed the technology scene, revealing different steps in its value chain and influencing investment trends. From the creation of models to infrastructure and application, every stage presents its own set of challenges, key players, and prospects.

- **Foundation model layer:** The bedrock of the generative AI ecosystem, this segment is dedicated to constructing and refining foundation models through rigorous research and development. Dominated by a handful of influential entities such as OpenAI, Anthropic, Google, and Meta, this space witnesses limited new entrants due to the steep challenges involved. Consequently, these pioneers hold substantial market influence.

- **AI infrastructure layer:** This segment involves a comprehensive suite of components encompassing data management, networking, hardware, and MLOps platforms, that supports the research, development, and deployment, and monitoring of AI projects. Hugging Face, with its emphasis on model and dataset collaboration, and Snorkel, a leader in training dataset management, are noteworthy contributors here.

- **AI application layer:** This segment revolves around tailoring and implementing models for specific applications. It benefits from the groundwork laid by upstream entities, merging domain knowledge with AI capabilities to drive commercialization. Despite the influx of players due to lower technical barriers, the competition remains fierce, often over limited market value slice.

**Industrial value chain**




Generative AI value chain
**Source:** Shixiang Research

The ascendancy of generative AI has catalyzed a surge in startup activity within the domain. In just the first half of 2023, funding directed towards this sector witnessed a five-fold increase compared to the entirety of 2022. As of the second quarter of 2023, the year is already setting records, with equity funding for generative AI startups exceeding $14.1 billion across 86 distinct deals.

- **AI Infrastructure:** The capital-intensive nature of crafting large language models has directed a major share of the funding towards the AI infrastructure segment. Since Q3 2022, this category has attracted more than 70% of the total funding, despite accounting for only 10% of all generative AI deals. Startups in their Series A funding round are, on average, securing ticket sizes exceeding $300 million. This substantial financial backing predominantly stems from investors' keen interest in foundation models, MLOps, and cutting-edge infrastructural technologies, such as vector database systems.

- **AI Application:** Since Q3 2022, approximately a quarter of the funding in the generative AI space has been channeled towards cross-industry generative AI applications and interfaces. Most of the activities are observed in text-centric applications, with content generation and chatbot assistants leading the charge.

As of November 2023, the generative AI sector boasts close to 1,500 enterprises, has witnessed 3,770 deals, and attracted more than $288 billion in investment capital. These startups span a wide spectrum of segments and modalities. However, a notable trend is that many application-layer companies are primarily focusing on tool-level solutions, rather than integrating with specific industry verticals.

Vertical-oriented generative AI models, tailored for specific industries, promise superior accuracy, enhanced operational performance, and heightened efficiency. Crafting these specialized models demands profound industry knowledge, which, in turn, elevates the technical challenges and barriers to entry.

02

# Essentials for building foundation models

**There are three essential resources that drive the successful execution of these stages: data, computing power, and talent.**

## 2.1 Data

To ensure the optimal development and performance of large language models, it is imperative for researchers to focus on three key aspects of **data:** scalability, quality, and diversity.

**Scalability** is paramount, given that these models are trained on vast amounts of unlabeled data in an unsupervised manner. This approach allows them to handle a plethora of downstream tasks. An optimal data-to-parameter ratio for training large language models stands at 20:1. This implies that training a robust large language model necessitates hundreds of billions of tokens.

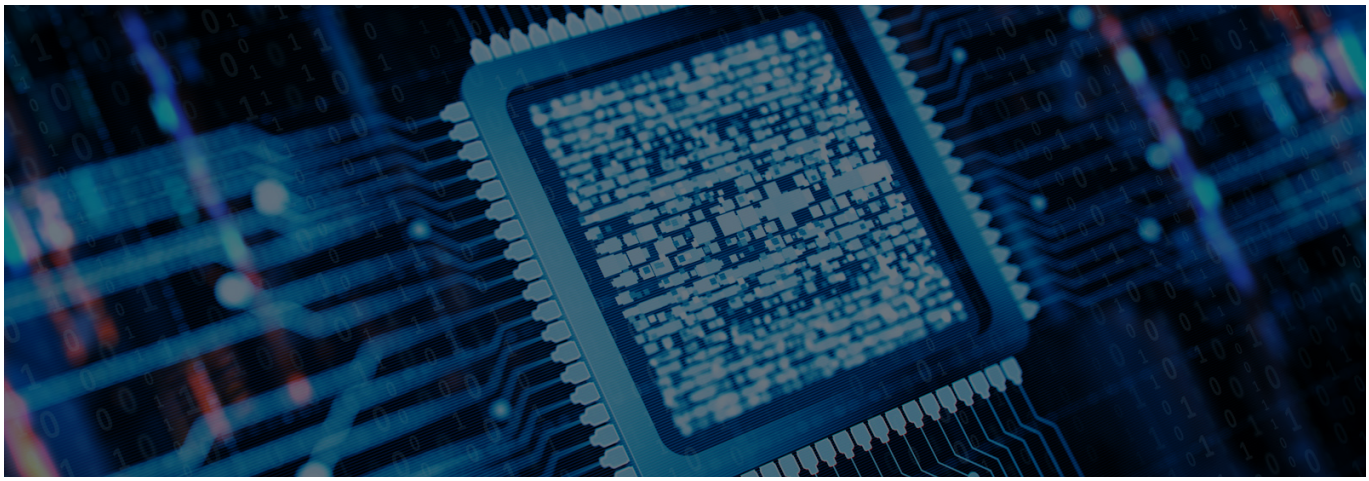
**Quality** is directly tied to the trustworthiness of a model. The adage “garbage in, garbage out” holds true; if the quality of the input data is subpar, the output will mirror that deficiency. Moreover, biases in training data will inevitably manifest in the model’s outputs. Given the sheer volume of data processed by large language models, ensuring quality becomes a monumental task. Engineers often resort to techniques such as hate and profanity filtering to discard harmful or inappropriate content.

On the other hand, **diversity** in data input plays a pivotal role in determining a model’s performance. Most large language models use a blend of diverse, public textual datasets. This can range from general content such as webpages, books, and conversations to more specialized datasets, such as multilingual content, scientific publications, and code.

The composition of a model’s training data is tailored to its intended purpose. For instance, models excelling in natural language processing tasks primarily draw from web content. However, versatile models might also integrate books, news, and conversational data to bolster their capabilities.

Those targeting a wider array of tasks, encompassing coding and scientific research, emphasize training with scientific data and code. It is crucial for researchers to curate meticulously the mix of data sources during the pre-training phase, as this composition profoundly influences the model’s efficacy across different tasks.

## 2.2 Computing power



**The creation of large-scale language models necessitates significant investments in both hardware and computational prowess.** Training for LLMs is typically executed on processors that are fine-tuned for matrix operations. As technology has advanced, each new generation of GPUs has brought with it an exponential increase in computing power:

- NVIDIA's GPU: Launched in 2022, the GPU H100 stands out as a versatile GPU, adept at managing a myriad of high-compute tasks in intricate settings. It has rapidly become a preferred choice for training expansive AI models. Moreover, NVIDIA is expected to release its new GH200 in 2024 that has the same computing power, but with triple the memory space.

- Cerebras' WSE-2: Tailored specifically for large-scale deep learning endeavors, the WSE-2 boasts the unique capability of hosting an entire model on a singular chip. This chip is optimized for sparse tensor operations, enabling it to execute vast amounts of parallel processing.

- Google's TPUv4: Introduced in 2021, the TPUv4 is Google's proprietary tensor processing unit, designed with TensorFlow computations in mind. It strikes an optimal balance between power and efficiency for extensive neural network training. Its seamless integration with Google Cloud services further cements its position as a go-to solution within Google's ecosystem.

Financially, training a model comparable to GPT-3 is no small feat. Training of OpenAI's GPT-3 took 34 days on 1,024 GPUs, with an estimated cost of $4.6 million in compute alone; Meta's LLaMa model consumed about 21 days on 2,048 A100 GPUs, roughly one million GPU hours, translating to costs exceeding $2.4 million. The subsequent phase, "inference" — where the trained model is used for predictions or text generation — can sometimes incur even higher expenses, especially for widely-used products such as ChatGPT.

However, there's a silver lining. Technological advancements have progressively reduced these costs. Nvidia, a titan in the AI chip market, remains at the forefront of this evolution, consistently churning out increasingly powerful chips tailored for large-scale models. Jensen Huang, Nvidia's CEO, is optimistic about the future, predicting that within a decade, AI will achieve a million-fold increase in efficiency, propelled by breakthroughs in chips, software, and other computing facets.

## 2.3 Talent

Developing a model on a par with GPT-3 is a resource-intensive endeavor, both in terms of human capital and financial outlay. Estimates suggest that a team of 70-80 specialists is required at various stages of the model's development.

It's worth noting that LLM development demands the crème de la crème of industry talent. Undoubtedly, top tech companies in generative AI, such as OpenAI, Google, and Meta, stand as the primary hubs for AI experts with profound knowledge and skills.

The network effect becomes a crucial factor when attempting to attract and retain such unparalleled industry talent. OpenAI's inception serves as a testament to the power of network effects in recruiting top-tier talent within the AI industry.

As the organization was taking shape, Greg Brockman, then CTO of OpenAI, sought the expertise of Yoshua Bengio, widely regarded as one of the pioneering figures in the deep learning movement. Their collaboration resulted in a curated list of the most eminent researchers in the domain.

By December 2015, Brockman had successfully onboarded nine of these luminaries, a cohort that included the likes of Ilya Sutskever, a former research scientist at Google Brain, who became Chief Scientist at OpenAI. The induction of such star researchers instilled a sense of accountability, generated publicity, and attracted talented individuals in the field. An illustrative example of this pull can be seen in the words of a Google employee, who joined OpenAI "partly because of a very strong group of people".

In addition to network effects, competitive compensation also helps attract top engineering talent. According to levels.fyi, OpenAI offers total compensation packages for engineering roles that can be 50% higher than peers like Google and Facebook.

03

# Future horizons of generative AI: scaling, alignment, and multimodality

**There are three essential resources that drive the successful execution of these stages: data, computing power, and talent.**

## 3.1 Larger models with stronger capabilities

Much larger foundation models are currently being developed by big players with more optimized data-parameter ratio. In 2020, OpenAI introduced the scaling laws that have since become instrumental in guiding the training of large models. These laws offer guidelines for striking the right balance between the size of an AI model and the volume of data used in its training. The overarching objective is to harness the available computational power in the most efficient manner. OpenAI's stance was clear: with an increase in computational resources, the emphasis should shift towards making larger models rather than inundating it with additional data.

As of June, GPT-4 stands out with its staggering one trillion parameters, more than five times bigger than GPT 3.5, while maintaining a 20:1 token to parameter ratio. In comparison, Google's PaLM2 boasts 340 billion parameters, while several other models hover around the 100 billion mark.

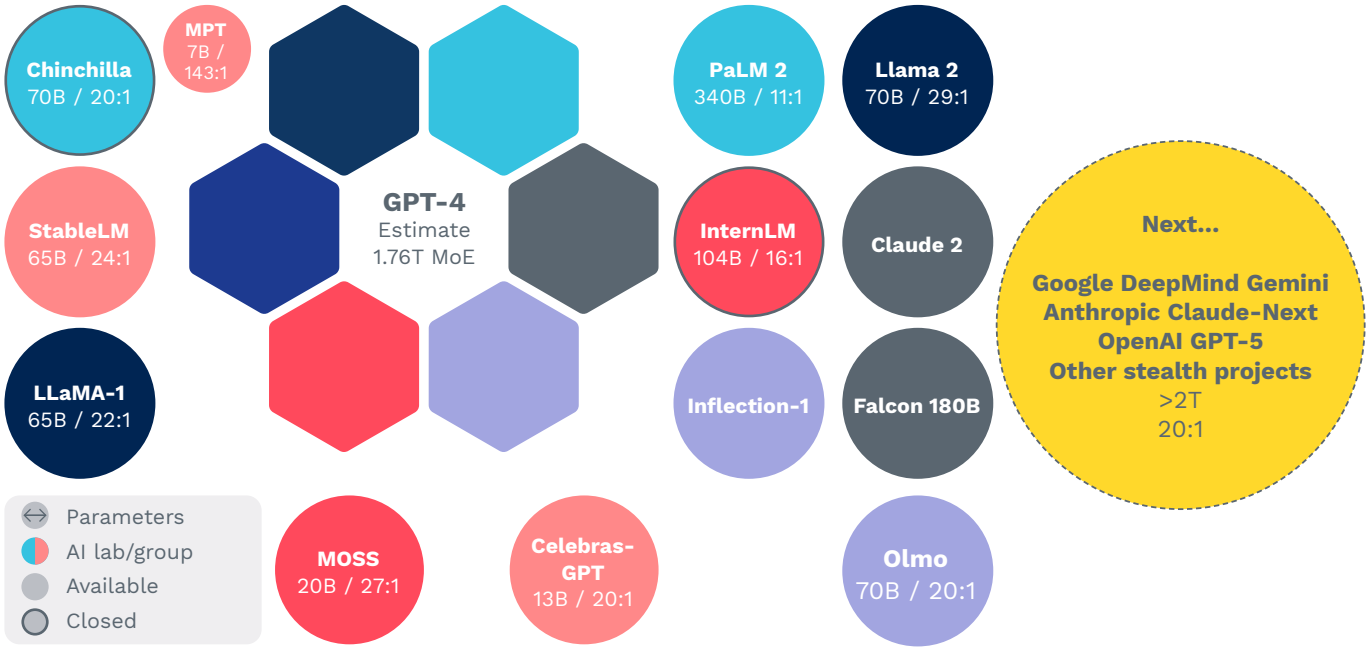
| **Model name**<br>Details | **AI Lab**<br>Openness |
|---|---|
| **GPT-4**<br>1T trained on 20T tokens* | **OpenAI**<br>API |
| **PaLM 2**<br>340B trained on 3.6T tokens* | **Google**<br>API |
| **PaLM 1**<br>540B trained on 0.8T tokens | **Google**<br>Closed |
| **Infection-1**<br>120B trained on 2T tokens* | **Inflection AI**<br>API |
| **InternLM**<br>104B trained on 1.6T tokens | **Shanghai AI**<br>Closed |
| **Chinchilla**<br>70B trained on 1.4T tokens | **DeepMind**<br>Closed |
| **Stable LM**<br>65B trained on 1.5T tokens | **Stability AI**<br>Open |
| **LLaMA-65B**<br>65B trained on 1.4T tokens | **Meta AI**<br>Open |

Existing top large language models by June 2023
**Source:** LifeArchitect.ai

The current large language models haven't yet approached the theoretical limits set by the scaling laws. Therefore, in line with the predictions of the scaling laws, industry giants are in the throes of developing even larger foundation models with more optimized data-parameter ratio. Anticipated models, such as Gemini, GPT-5, and Claude-Next, are projected to boast around two trillion parameters, adhering to the 20:1 token to parameter ratio.

Yet, it›s essential to temper expectations. While the allure of exponential growth is tantalizing, practical limitations, primarily hardware constraints, mean that the next generation of models will likely be only two to three times larger than their current counterparts.

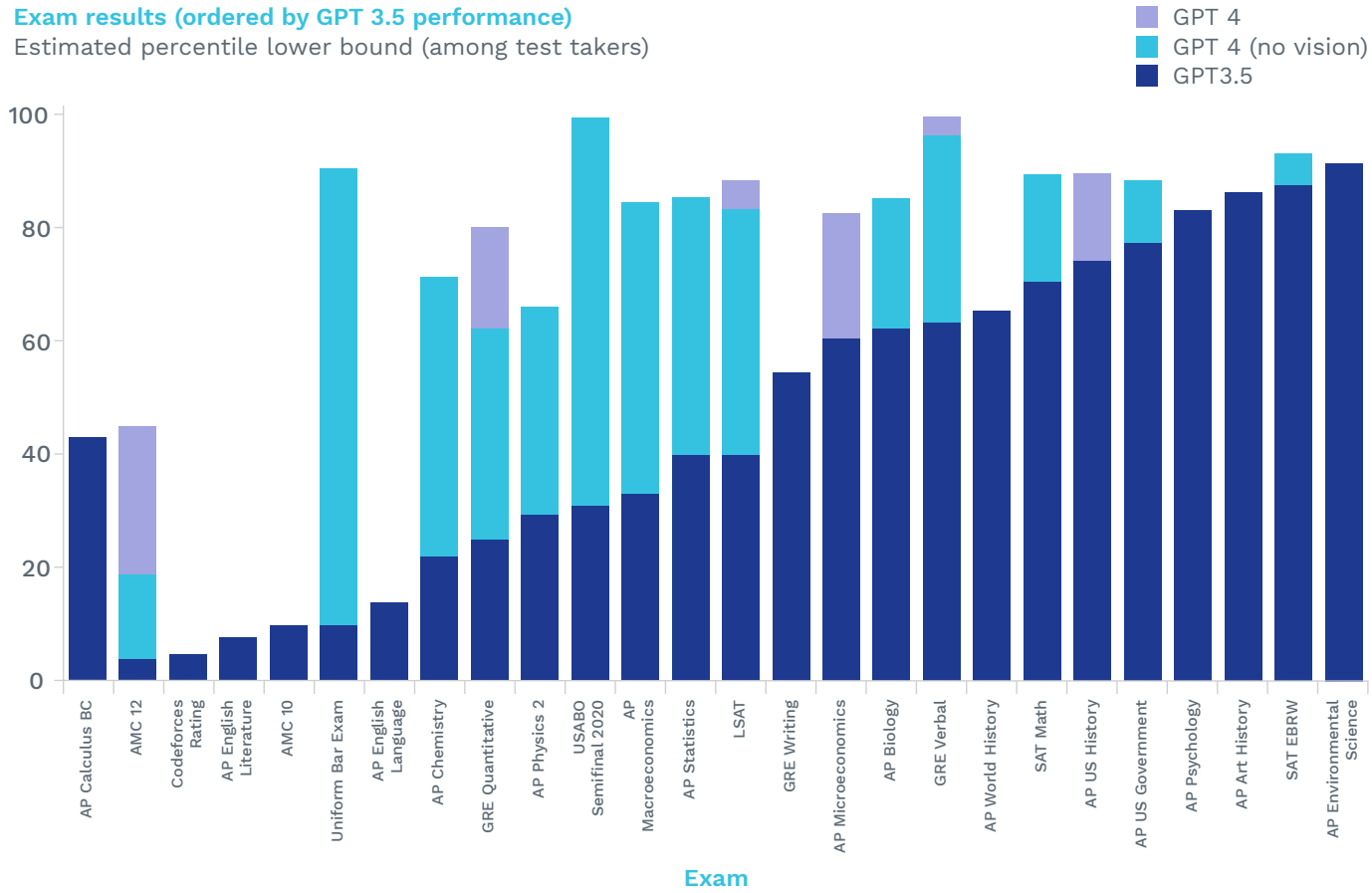
## 2023-2024 optimal language models




2023-2024 optimal language models
**Source:** LifeArchitect.ai

## Exam results (ordered by GPT 3.5 performance)

Estimated percentile lower bound (among test takers)



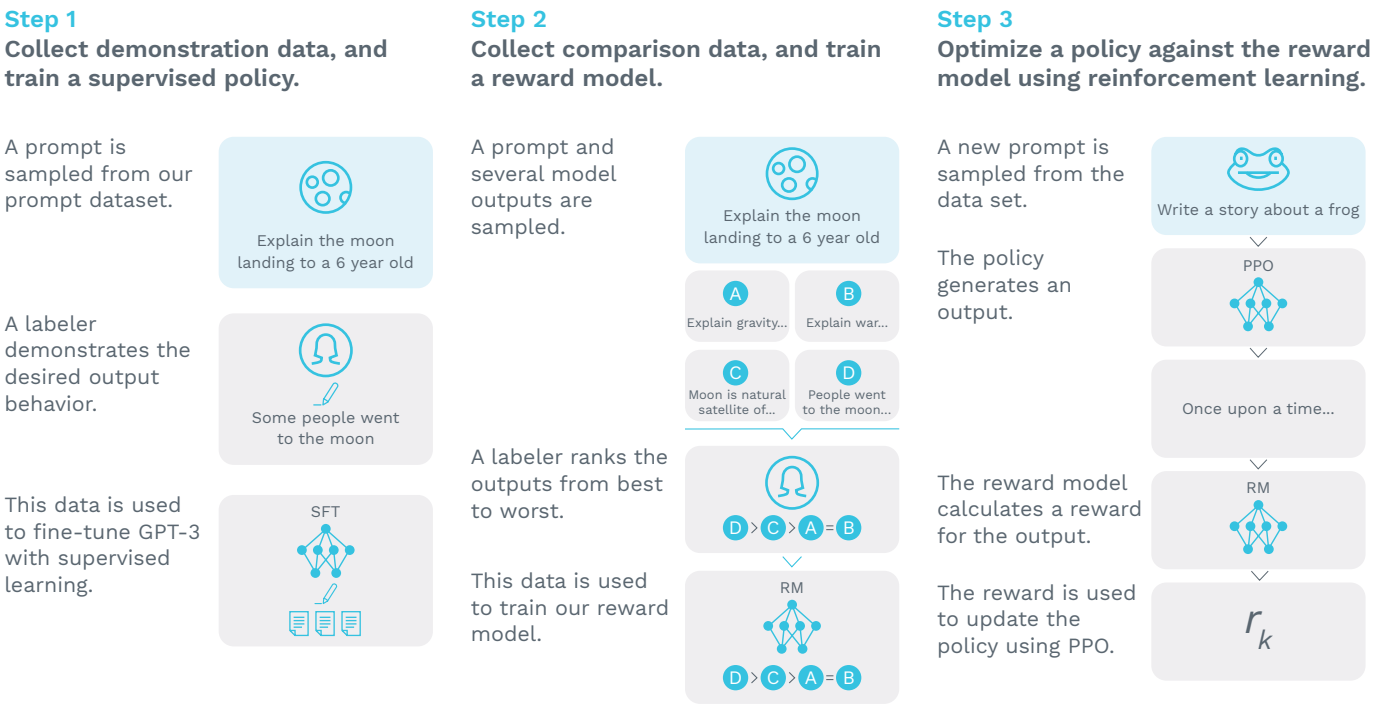

**Source:** OpenAI

## 3.2 AI alignment

With the rise of large AI models, we've witnessed an influx of advanced capabilities becoming cornerstones in socio-economic development. Predictions suggest that these models could add up to $4.4 trillion to the global economy annually. Yet, a significant challenge emerges: aligning these models with human values and ethics.

This "value alignment" is vital, as larger models can pose higher risks – especially since they're often trained on internet data, inheriting its biases. Unaligned models risk perpetuating discrimination or assisting in harmful activities. Ensuring safe outputs and preventing misuse is, thus, a crucial focus in today's AI alignment efforts.

In the current landscape, AI alignment predominantly leverages reinforcement learning from human feedback (RLHF). In essence, a model is trained to resonate with human intentions and subsequently undergoes reinforcement learning to adjust the large language model. OpenAI, in its pursuit of alignment, also employs methods such as training auxiliary models to aid humans in evaluating complex tasks and developing systems that can expedite alignment research beyond human capabilities.

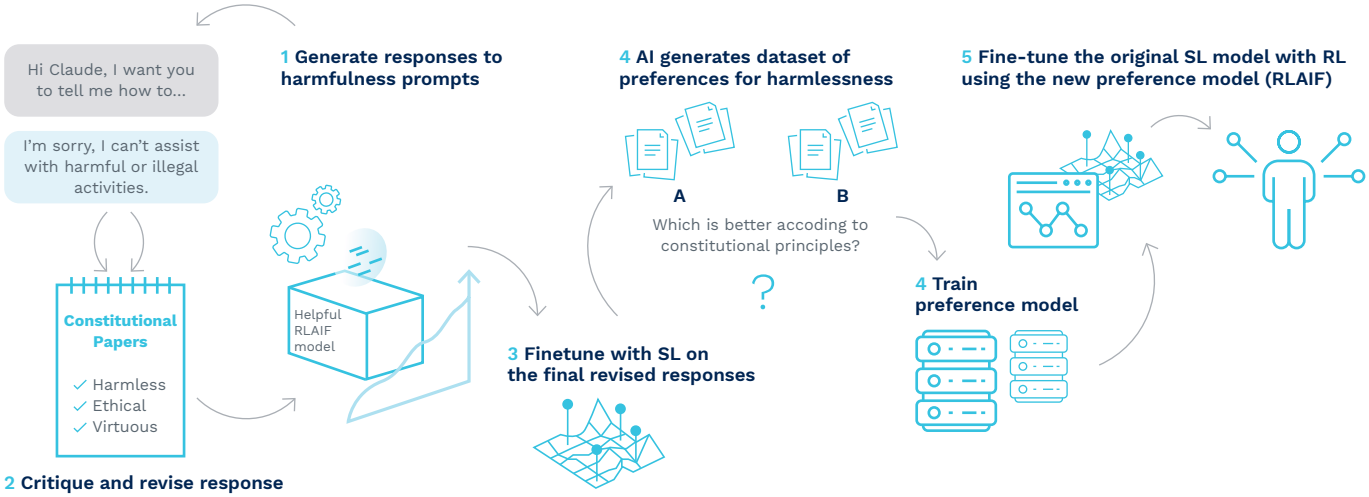



Reinforcement Learning from Human Feedback used for GPT training
**Source:** OpenAI

The “constitutional AI” approach, championed by US-based AI firm Anthropic, shifts from “human supervision” to scalable oversight. Instead of solely relying on humans, this strategy uses a subordinate AI model to ensure that the primary model’s outputs align with set “constitutional” principles. Drawing from diverse standards, including the Universal Declaration of Human Rights, Anthropic’s model, Claude, effectively minimizes harmful content while delivering valuable outputs – showcasing a a balance of usefulness, transparency, and safety in AI systems.

**Supervised Learning (SL) Stage**
Collect demonstration data, and train a supervised policy.

**Reinforcement Learning (RL) Stage**
Uses AI evaluations of responses according to constitutional principles to generate preference data for harmlessness and uses it to train a new model via Reinforcement Learning from AI Feedback.




Constitutional AI
**Source:** Anthropic

In July 2023, OpenAI announced the establishment of a specialized AI alignment team, designated as “superalignment.” This team’s core mission is to discern methods to guarantee value alignment and safety in next-generation AI systems. To bolster this mission, OpenAI is allocating 20% of its computational power towards this cause. Central to this initiative is the innovative approach of employing AI technologies to assist humans in addressing the critical issue of AI value alignment.

To harness the full potential of AI, it is vital that the technology’s objectives resonate closely with human values and intentions. Achieving this harmony requires a collaborative approach, drawing from various disciplines and sectors. Government bodies, industry leaders, and academia should collaboratively invest resources to align AI’s values, ensuring that our capability to supervise, understand, and control AI’s evolution keeps pace with its rapid advancements – ultimately benefiting humanity as a whole.

## 3.3 Multi-modality

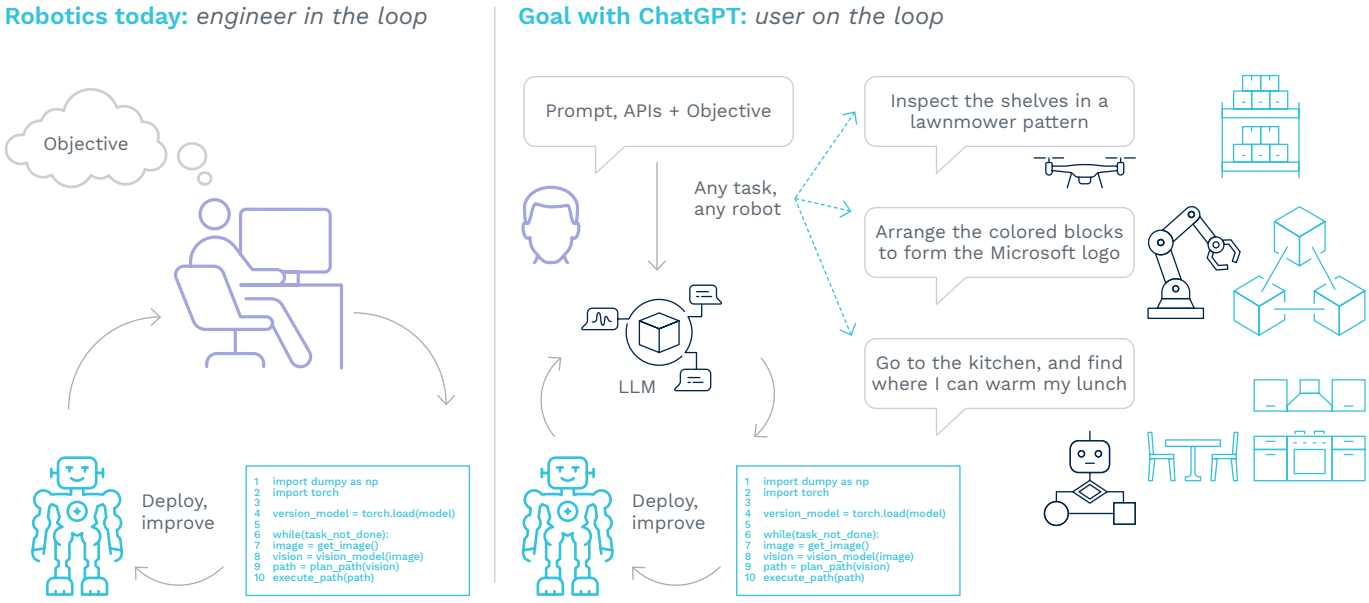
Multimodal AI represents a significant evolution in the realm of artificial intelligence. By harnessing information from diverse modalities e.g., text, images, audio, video, biodata, or sensory data – these applications can comprehend and generate outputs across a spectrum of data types. The power of multimodality lies in its ability to tap into both complementary and redundant data streams. Consider a generative AI model trained on text and images: it could produce a textual description for a given image or, conversely, generate a visual representation based on textual input.

Multimodal AI has applications in areas such as robotics, self-driving cars, multimedia search tools, and advanced virtual assistants. The convergence of foundational models with robotics is particularly gaining momentum. Big names and new startups alike are jumping on this trend, signaling a future where AI interacts directly with our physical environment.

### 1. Microsoft and ChatGPT

Microsoft is at the forefront of integrating large language models with robotics, harnessing the capabilities of ChatGPT to redefine human-robot interactions. In groundbreaking research conducted by the Microsoft Autonomous Systems and Robotics Research Group, ChatGPT has been instrumental in enabling intuitive control over a diverse range of robotic platforms, from drones to home assistant robots, using natural language commands.

This innovative approach merges language comprehension with physical actions, bypassing the traditional need for manually coded robotic controls, and heralding a future where robots understand and act upon human language seamlessly.

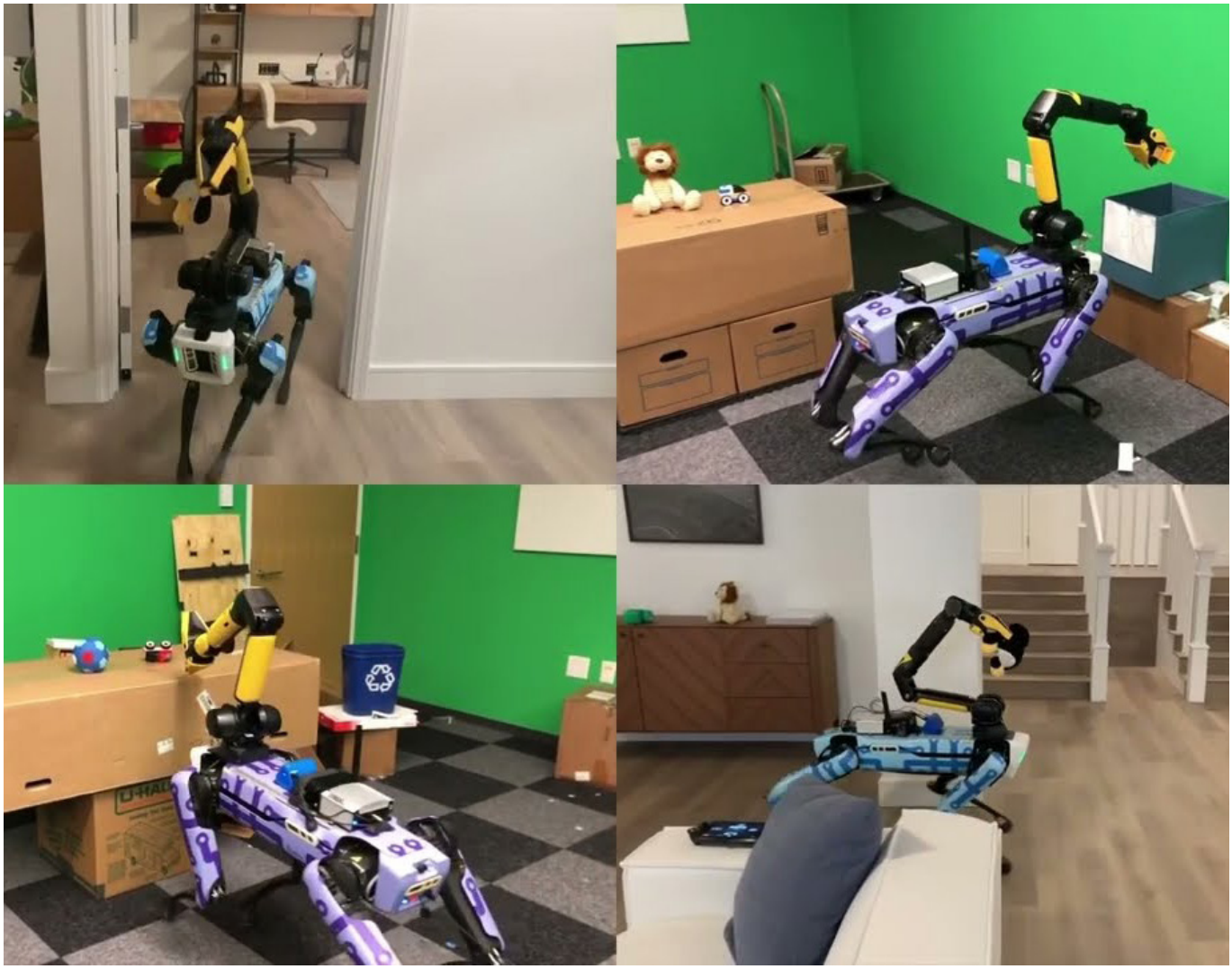



ChatGPT frees the engineer from the loop and engages users to provide feedback
**Source:** Microsoft

**2. Meta and Spot**

Meta, in partnership with Facebook AI Research, is pushing the boundaries of robotics by integrating large language models with Boston Dynamics' renowned Spot robot. Their research is centered on amplifying Spot's cognitive abilities, allowing it to understand and act upon simple natural language directives.

This enhancement empowers Spot to navigate unfamiliar terrains adeptly. Key advancements include the inception of the artificial visual cortex VC-1 and the pioneering adaptive sensorimotor skill coordination (ASC) technique. These innovations have notably augmented Spot's proficiency in locating and retrieving objects across diverse environments, signifying a leap towards the realization of universally intelligent embodied AI agents.

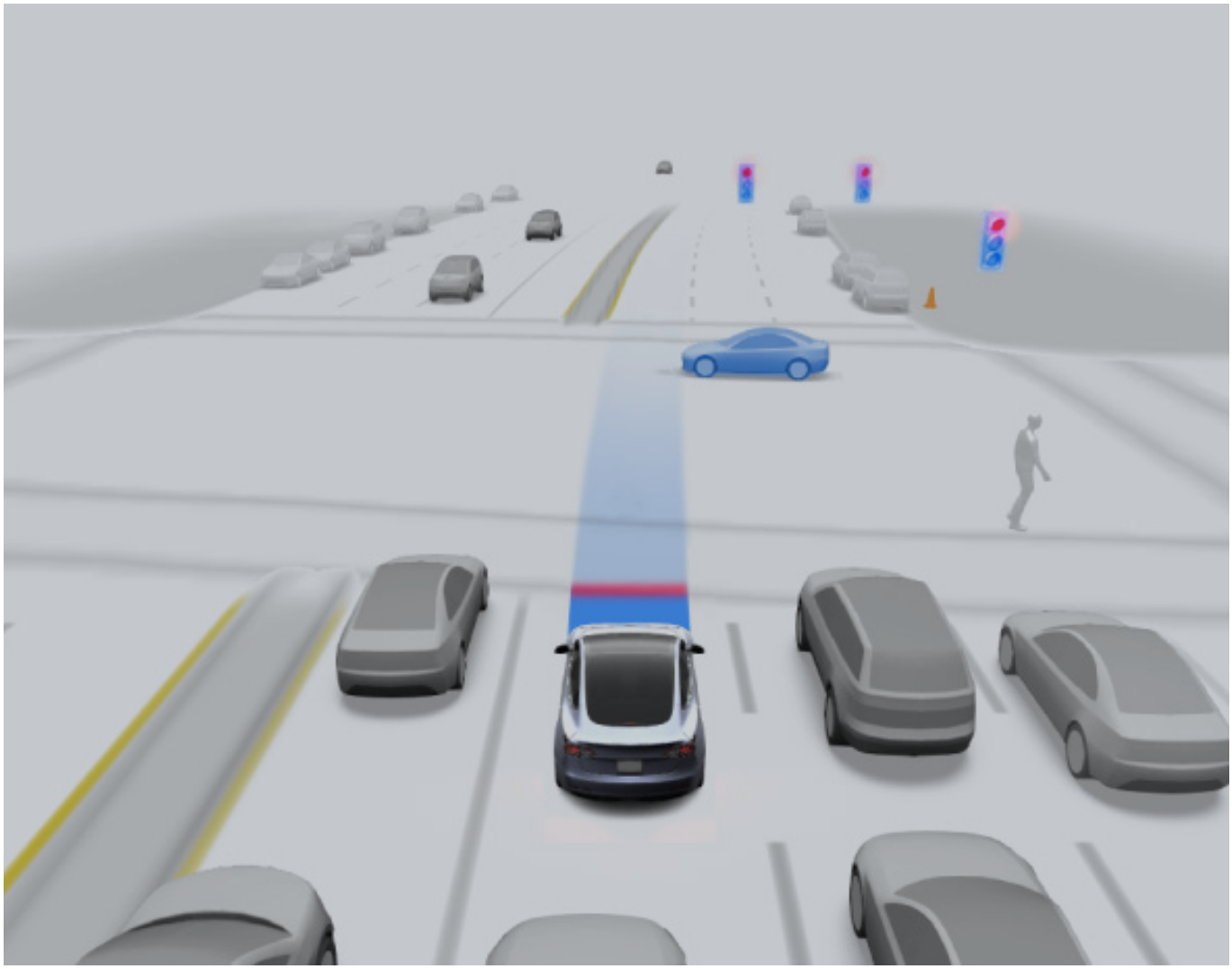


**3. Tesla FSC V12**

Tesla's full self-driving software version 12 (FSD V12) marks a pivotal leap in semi-autonomous driving technology. Designed to emulate the human brain, the system harnesses the power of neural networks and utilizes eight cameras, capturing at a rate of 36 frames per second, to navigate and make decisions. Rather than relying on explicit coding for road features, FSD V12 learns from videos sourced from millions of Teslas on the road, allowing it to emulate human driving behaviors.

This innovative approach is a departure from its predecessor, V11, which had over 300,000 lines of C++ for explicit controls. With test drives being conducted globally, including in countries such as New Zealand and Norway, Tesla's FSD V12 showcases the brand's commitment to advancing artificial intelligence-driven autopilot capabilities.

Certainly, while progress in multimodal AI has been swift, we are only at the beginning of this expedition. As of now, when confronted with complex situations, multimodal models have not matched the comprehensive prowess demonstrated by linguistic-focused models such as GPT. Yet, the speed of innovation is staggering. Given the momentum, we can anticipate significant breakthroughs in the near future.



# Conclusion

**The emergence of generative AI is more than just a technological advancement; it's a paradigm shift in how we perceive and interact with machines. As foundation models continue to shape industries, drive innovation, and redefine the boundaries of what is possible, it becomes imperative for us to understand, adapt, and harness their potential responsibly.**

**The future of generative AI promises not only enhanced capabilities, but also new challenges. Nevertheless, one thing is certain: the journey of generative AI is just beginning, and its impact on business and society will be profound and lasting.**

Published by:

**Mohamed bin Zayed**
**University of Artificial Intelligence**
Masdar City
Abu Dhabi
United Arab Emirates

**mbzuai.ac.ae**
**mbzuai.ac.ae/llm/**
**mbzuai.ac.ae/institute-of-foundation-models/**