Contribution to Generative AI market analysis

Technology and market trends

Version:	1.2
Date:	May 2024
Author:	Fournigault, Mike

European Commission

European Commission

Contents

Glossary

This glossary is inspired by the glossary of [DR10] and completed by the authors of this report.

Application programming interface (API) is a way to programmatically access (usually external) models, data sets, or other pieces of software.

Artificial intelligence (AI) is the ability of software to perform tasks that traditionally require human intelligence.

Artificial neural networks (ANNs) are composed of interconnected layers of software-based calculators known as "neurons." These networks can absorb vast amounts of input data and process that data through multiple layers that extract and learn the data's features.

Deep learning is a subset of machine learning that uses deep neural networks, which are layers of connected "neurons" whose connections have parameters or weights that can be trained. It is especially effective at learning from unstructured data such as images, text, and audio.

Foundation models (FM) are deep learning models trained on vast quantities of unstructured, unlabeled data that can be used for a wide range of tasks out of the box or adapted to specific tasks through fine-tuning. Examples of these models are GPT-4, PaLM, DALL·E 2, and Stable Diffusion.

Fine-tuning is the process of adapting a pretrained foundation model to perform better in a specific task. This entails a relatively short period of training on a labeled data set, which is much smaller than the data set the model was initially trained on. This additional training allows the model to learn and adapt to the nuances, terminology, and specific patterns found in the smaller data set.

Generative AI is AI that is typically built using foundation models and has capabilities that earlier AI did not have, such as the ability to generate content. Foundation models can also be used for nongenerative purposes (for example, classifying user sentiment as negative or positive based on call transcripts) while offering significant improvement over earlier models. For simplicity, when we refer to generative AI in this report, we include all foundation model use cases.

Discriminative AI is the AI prior to generative AI. Tasks performed by discriminative AI are usually much simpler than the ones performed by generative AI. Discriminative AI typically uses models that classify or predict based on existing data.

Backbone architecture defines the structure of the deep neural network used in models. For foundation models, the backbone architecture is usually very large and complex incorporating specific blocks with dedicated mathematical functionalities.

Graphics processing units (GPUs) are computer chips that were originally developed for producing computer graphics (such as for video games) and are also useful to accelerate computations of deep learning, notably during the learning phase. Now semi-conductors' founders like Nvidia or AMD design specific GPUs only for AI.

Tensor processing units (TPUs) are computer chips that are designed by Google to accelerate computations of deep learning. On the contrary of GPUs, TPUs have always been dedicated exclusively for deep learning computing.

Large language models (LLMs) make up a class of foundation models that can process massive amounts of unstructured text and learn the relationships between words or portions of words, known as tokens. This enables LLMs to generate natural-language text, performing tasks such as summarization or knowledge extraction. GPT-4 (which underlies ChatGPT) and LaMDA (the model behind Bard) are examples of LLMs.

AI-generated content (AIGC): some specialists prefer to use the term AIGC rather than generative AI because some backbone architectures like Transformers are used in foundation models but also in discriminative models that don't generate any content.

1. Executive Summary

Fujitsu is pleased to submit its input on the topic of generative AI systems in response to the European Commission call for contributions which aims at gathering specific information and views in relation to competition aspects on the subject.

As a global leader in AI, Fujitsu has been promoting the research and development of innovative AI and machine learning technologies for over 30 years, and continues to embrace recent breakthroughs in areas like generative AI. To date, Fujitsu boasts a track record of more than 7,000 AI customer use cases in fields including manufacturing, retail, healthcare, public safety and more.

Generative AI is increasingly seen as a game changer for business. Over the last few months, the arrival of more powerful and capable Generative AI, such as OpenAI's GPT, has driven an explosion of interest in the potential of this exciting technology to make organizations more competitive and enhance the services they deliver.

In this short paper, we will share some of the key points leading to our analysis on competition in the Generative AI field, starting with a brief introduction, its history, and its core Architecture.

With the description of the architecture, we intend to show that :

- Many components are required in the technology stack to develop, provide, or integrate generative AI models.
- Large scale of models became a requirement for performance with many consequences on the technical requirements and costs of generative AI models.

These two statements have major impacts on the market structure.

We will also describe how data plays a pivotal role in generative AI, breaking down its significance and key characteristics.

The next chapter will conclude our analysis of the Generative AI market and competition, highlighting:

- The drivers of Competition
- The barriers to Entry and Expansion
- The emerging competition issues
- The difference between Open source versus proprietary Generative AI
- The monetization strategy
- The market Consolidation

2. Introduction of Generative AI & its core Architecture

2.1 A brief introduction on Generative and its application

Generative AI is a subfield of artificial intelligence focused on creating new content, rather than analyzing existing data. This "creation" can encompass text, images, audio, and more, with models learning patterns and relationships within datasets to generate similar but novel outputs.

Some of its applications includes:

- By task:
  - For text: translation, summarisation, recognizing intents and entities of text, composing reports (from meeting notes for example), writing program codes, composing emails, letters or poems in different styles. In can also be to describe an image or a video, to translate voice to text, or to extract knowledge.
  - For image or video: image restoration, image edition by removing or adding objects, transforming an image into a painting with a specific style. It can be also to generate images or videos, realistic or not, from a simple text description.
  - Sound: generate speech from text, transform a speech from a given voice into another one, generating music from a text description, ...
  - Graph: generate knowledge and relation graphs from text, images, videos, ...
  - 3D: generate 3D models from text description, transforming 3D models by following specific styles.
  - generating contents like realistic images and videos, composing music, writing different kinds of creative text formats like poems, musical pieces, email, letters, data analysis, etc.
- By Industry:
  - In software development: Generating code, code documentation, test automation.
  - In game industry: generating code, generating 3D models, generating game scenarios, interactively renewing the game content with user interactions, generating music.
  - In movie industry: generate scenarios, supporting all the computer-based content creation, performance of passed actors, generating subtitles, restoring old movies.
  - In drug discovery: Designing new molecules with desirable properties.
  - In product design: Developing prototypes and variations quickly.
  - In manufacturing: production planning, quality control, etc.

The Figure 1 gives a more complete overview of Generative AI applications by tasks and by industries.

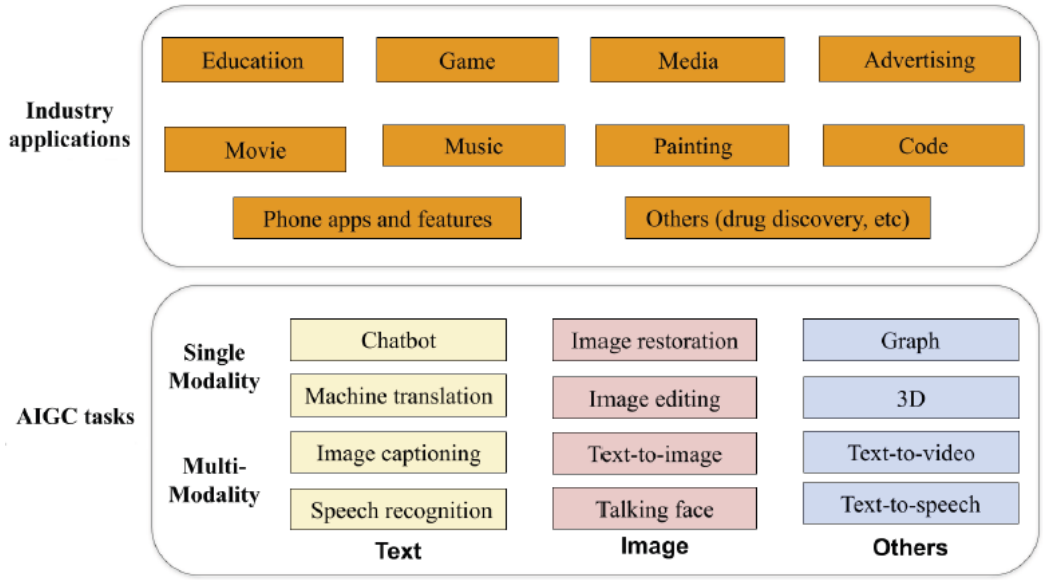


Figure 1 Overview of generative AI: core AIGC tasks and industrial applications [DR6].

Generative AI primarily relies on generative models, which learn the underlying statistical distributions of a given dataset. Unlike discriminative models that classify or predict based on existing data, generative models can produce entirely new content that adheres to the learned patterns. For example, where a discriminative model can state whether if a painting is made by Van Gogh, a generative model can transform an image into a painting following the style of Van Gogh.

By the time, Generative AI Models have been able to provide unprecedent capacities to its users:

- The ability to generate realistic content without having any expertise in AI,
- The capacity to automate the creation of large amounts of content in a short amount of time.

2.2 History of Generative AI and its sudden success with the public

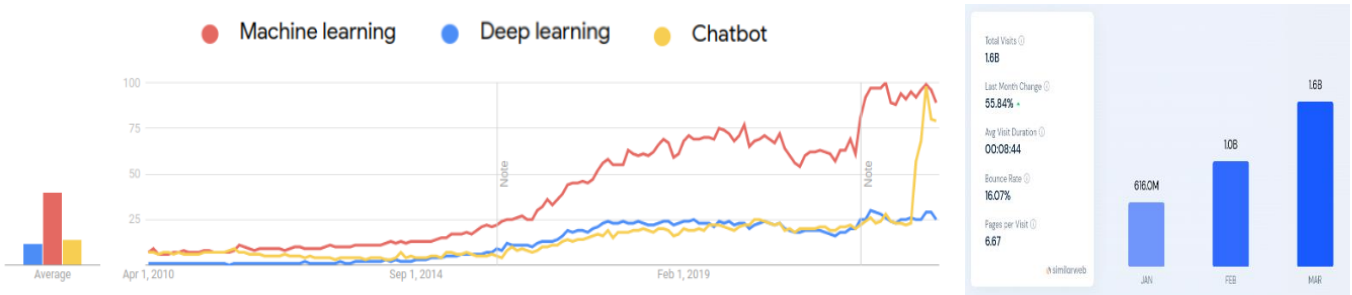
The history of Generative AI is surprising in more than one way.

No significative results were made since the first tries in the 1950s. **2013 and 2014 have been a significant milestone for Generative AI** with the creation of Variational Autoencoders (VAE) and Generative Adversarial Networks (GANs) [DR3], due to their impressive results in various applications. For example, GANs were able to generate realistic-looking faces but entirely fictitious.

In parallel, progress were made in the language domain, until the intersection emerged in 2017 with the creation of the transformer architecture. The success of transformer models has led to the development of large-scale, pretrained models, for language such as OpenAI's GTPs or Google's BERT, for vision such as ViT, for mixing language and vision such DALL-E (for example to generate an AI from a text prompt).

2022 and 2023 have been breakout years for AI with the adoption of Generative AI by the general public, i.e., non-expert users, creating an unprecedent acceleration in AI democratization.

In 2022, OpenAI made its latest iteration of the Large Language Model, GPT-3, accessible to the public through the introduction of ChatGPT—an interactive chatbot interface. This platform enabled individuals to issue instructions to the model, marking a significant milestone in democratizing access to advanced language processing technology. The disruption was so important that the interest and adoption by the public exploded with a ramp-up of 100 million monthly active users after only two months, a worldwide record (4.5X faster than Tiktok) [DR4]. In March 2023, ChatGPT had 1.6 billion monthly active users.



The authors of [DR6] argue that the factors contributing to the advent of such powerful tools can be summarized from two perspectives: content need and technology conditions.

From the perspective of content need, the Internet has transformed communication and interaction with the world, particularly through the evolution of digital content. In the Web 1.0 era, characterized by static websites and one-way communication, Professional Generated Content (PGC) dominated, created by professionals like journalists. The transition to Web 2.0 saw the rise of User Generated Content (UGC) on social media platforms, with increased volume but potentially lower quality. As we move towards Web 3.0, characterized by decentralization, intermediary-free structures, and the need to balance volume and quality, artificial intelligence (AI) is recognized as a promising tool for generating diverse content types, such as images, at an unprecedented speed and quality for non-professionals. The shift towards AI Generated Content (AIGC) has the potential to greatly transform the way data is consumed and produced, giving individuals and organizations more control and flexibility in the content they create and consume [DR6].

The models of Generative AI are deep learning models. The performance of deep learning models, particularly in generative AI tasks, relies heavily on the size of the training dataset. Advances in data access, facilitated by the

Internet, have enabled the utilization of massive datasets, while improvements in computing resources, especially the shift from CPU (Central Processing Unit, the processor controlling every computer or smartphone) to AI accelerators like GPUs (Graphic Processing Unit) and TPUs (Tensor Processing Unit), contribute significantly to the development of AIGC by providing the necessary power for training large deep learning models at more affordable prices.

Major technology corporations such as Microsoft, Google, and Apple, along with influential entities like OpenAI and Anthropic, are currently engaged in a competitive endeavor to advance the development of cutting-edge models. They aim to offer the general public and AI application developers access to remarkable capabilities. A comprehensive analysis of the release dates of Large Language Models [DR5] reveals the ongoing and dynamic competition within this domain.

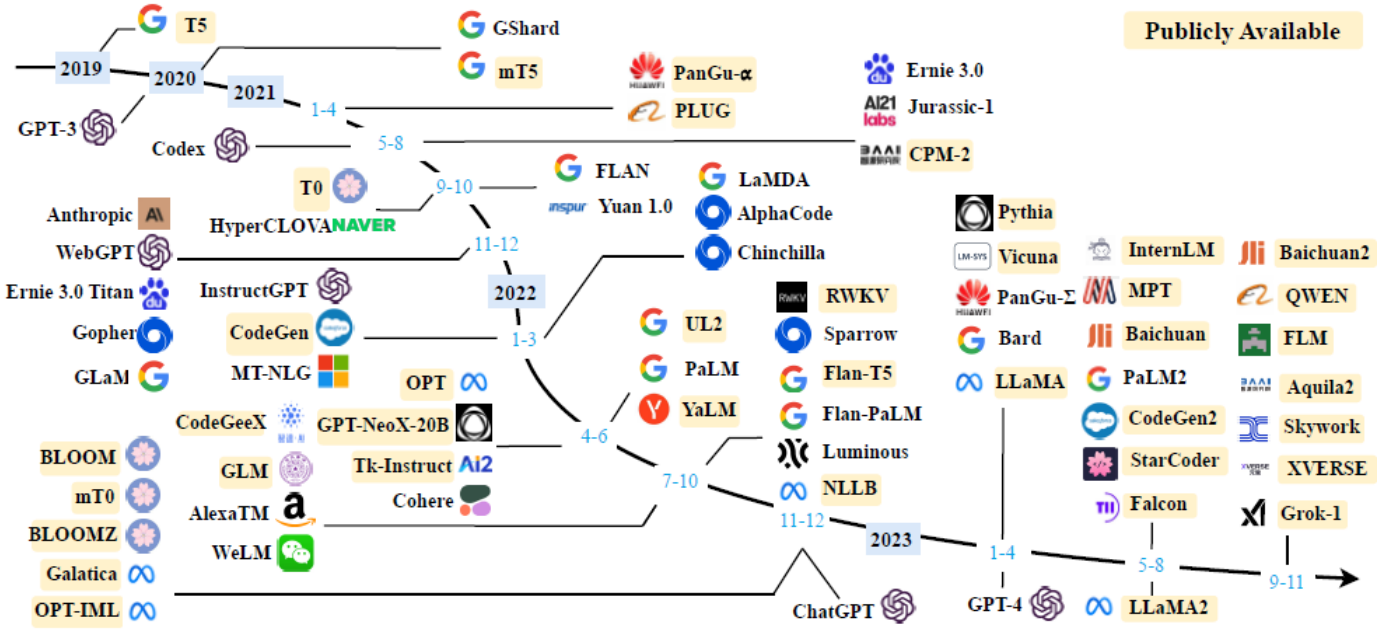


Figure 2 Overview of Large Language Models releases by the major actors of the market.

In terms of revenues, the market of Generative AI represented 24 billion of dollars in 2023, and Gartner projects an increase to 60 billion of dollars by 2026.

Before going deeper in the market analysis, we think that it is important to highlight some of the fundamental characteristics of Generative AI, to better understand their impacts on the market:

- Generative AI roots in deep learning advancements, especially neural networks like Variational Autoencoders (VAE) or Generative Adversarial Networks (GANs). As all deep learning technics, it takes advantage of increasing big volume of data to increase its performance, it also requires a significative computing power to train the models.
- Generative AI models are probabilistic by nature, and not deterministic. They estimate a probability to observe a given sample, this is foundational of their generation capacities. But however, the realism of the generated content is they are inherently prone to errors.
- Generative AI models are much more complex than their discriminative brothers. They are deeper (more layers included in the network structure), they are wider (include specialized substructures like self-attention blocks). As an example, one of the most complex discriminative models working on images has 63 million of parameters, where the first version of the model Dall-E, generating images, has 12 billion of parameters. Their complexity enables them to perform AI tasks that are much more complex than the ones processed by discriminative models. Their complexity requires also much more computing power to train the models, but also much more efforts and talents to train the models.

2.3 Core architecture of Generative AI systems

In this section, we want to give an overview of the Generative AI technology stack, to identify the key components and the key players of the market. This overview does not intend to be exhaustive but representative of the most dynamic actors of the market.

In relation to any artificial intelligence (AI) system, the composition of its components is contingent upon the perspective of the actor assessing the system. We propose a comprehensive examination of these components in the context of distinct actors:

- AI Model Developer: This individual is responsible for defining the model's architecture, conducting the model training, and evaluating its performance.
- AI Model Provider: The provider's role involves preparing and packaging the AI model for deployment, whether hosted in the cloud or embedded in a device.
- AI Model Integrator: Integrators utilize the AI model within their services or applications to augment them with intelligent capabilities.
- End-User of the AI System: The end-user leverages the AI system, provided by the integrator, to assist in performing tasks efficiently.

It is essential to acknowledge that an actor may assume multiple roles. Numerous AI model developers offer access to their models as web services through Application Programming Interfaces (APIs), exemplified by entities such as OpenAI and Stability.ai. Furthermore, some developers also function as integrators, providing end-user applications, as observed in the cases of OpenAI's ChatGPT and Google's Gemini assistant.

For readers not interested in understanding all the architecture details of Generative AI, we provide the section "2.3.1Key takeaways of the core architecture".

For readers interested in understanding all the architecture details of Generative AI and how they impact the market, we advise to continue in section "2.3.2 Core Architecture for the model development"

2.3.1 Key takeaways of the core architecture

Core architecture for the model development

- Data is fundamental: Training requires vast amounts of diverse data, typically several terabytes, impacting model capacities and performance and posing challenges in collection, storage, and pre-processing.
- Generative models: These models (e.g., GPT-3, Stable Diffusion) are built on specific backbone architectures (Transformers, GANs) with billions of parameters and trained in several steps using various techniques (self-supervised, fine-tuning, reinforcement with feedback). The scale of training is a key requirement for performance and during the training, each entire dataset is ingested dozens of times to optimize the neural network.
- Model complexity: Models can be unimodal (single input/output type) or multimodal (handle diverse data types), with increasing complexity leading to higher demands on training and infrastructure.
- Training and infrastructure: Training requires large team with significant expertise, large computing clusters with AI accelerators like GPUs, and a significant software stack (frameworks, libraries). For example, Meta trained its first version of LLAMA on a cluster of 2048 high-end GPUs for 21 days. Computation costs for a training run of models like GPT-3 or LLAMA2 are estimated of several million dollars, multiple runs are necessary to release a product-grade model.
- Barriers to entry: The need for very large datasets, complex models, and specialized knowledge creates significant barriers for new entrants in the generative AI field.

Core architecture for the model provision

- Access and Deployment Options: Models can be used for pre-built applications or customized, deployed in various environments (provider's one, public cloud, private cloud) with varying costs and control levels.
- Model-as-a-Service (MaaS): is the standard approach chosen by model developers to offer API access and hide the architecture complexity and infrastructure costs for users.

- Specialized AI Platforms: Platforms like Fujitsu Kozuchi provide cloud abstraction, but also additional features, and security measures to mitigate risks of biases, fairness and hallucinations of foundation models that remain after training.
- Private Deployment: If the model is distributed by the developer as file artifacts, an option for higher control is to deploy it in a private environment but it requires building and managing the infrastructure.
- Scalable Infrastructure: regardless of deployment configuration chosen for model serving, each model provider must leverage large GPU clusters scaling up or down with data volumes and number of users. It may cost at least 38 million dollars per month to OpenAI to host the model of ChatGPT for 1 million active users per hour.

Core architecture for the model integration

- Integration Options: Models can be integrated into dedicated Generative AI applications (like ChatGPT) or as features in existing applications like Copilot for Microsoft 365 or Copilot for Github.
- Out-of-the-Box vs. Custom Applications: Pre-built applications exist for specific tasks like chat or image generation, but custom applications can be built for diverse use cases.
- Developer Preference: A survey driven by Altman Solomon suggest that 94% of application developers plan to integrate Generative AI based on off-the-shelf models, where 70% of application developers prefer training and deploying their own models for discriminative AI.
- Simplified Architecture: Thanks to MaaS, integration requires less complex architecture compared to development and provision, and focuses on data interfaces, pipelines, and ethical considerations.
- Key Components: Integration utilizes APIs for communication with models, leverages AI ethics tools for fairness and accuracy, and emphasizes monitoring, security, and logging for optimal performance.

2.3.2 Core Architecture for the model development

The core architecture for the model development is structured with the following layers:

- The **data layer**: it is the lifeblood of any generative AI model. Data required to train the model are collected, pre-processed, labelled and updated.
- The **generative model layer**: it plays a critical role in shaping the model's capabilities and outputs. In this layer, the model architecture is defined, the training algorithms are developed, one or several versions of the model are trained and evaluated. Maintenance procedures are also defined to fix and update the models to keep improving their performances. Other considerations such as interpretability and controllability, efficiency and stability, or domain-specific adaptations are usually parts of this layer as well.
- The **infrastructure layer**: Training a generative AI model requires a specific set of infrastructure characteristics such as: high performance computing with massive parallel processing and fast interconnection between processors, high-speed storage and efficient data pipelines, software and frameworks to develop, manage and orchestrate the training jobs.

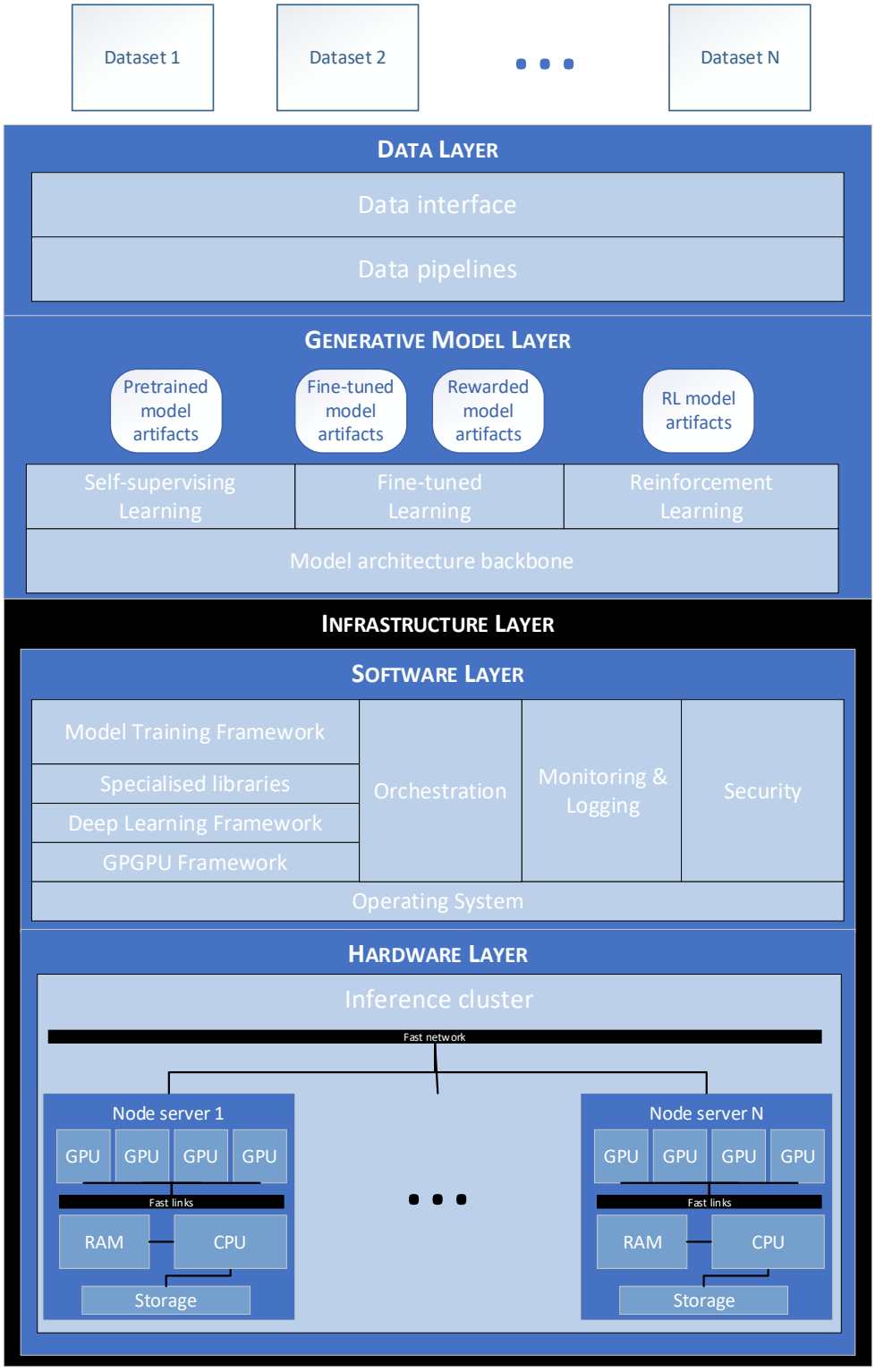


Figure 3 Core architecture for the model development

The data layer is the foundation of the model development.

Data types include text, images, audio, video, code, etc., depending on the desired output format. Text generation models utilize written documents, while image generation requires imagery datasets.

To train their baseline, generative AI models require more data than any of their predecessors. According to publicly available data [DR5]:

- In May 2020, the model GPT3 (by OpenAI) required a dataset of 300 billion of tokens[1] for its training.
- In June 2020, the model G-Shard (by Google) required a dataset of 1 trillion of tokens for its training.
- In July 2023, the model LLAMA2 (by Meta) required a dataset of 2 trillion of tokens for its training.
- In July 2023, the model Stable Diffusion XL (by Stabiliy.ai) required a dataset of 2.3 billion of image-text pairs for its training.

Three hundred billion of tokens if equivalent to 160000 hours of human reading, such volumes of data require to dispose of a public dataset pre-collected, or to tremendous efforts to collect these data.

The data sources need to be representative of the domain where to apply the model, non-biased and inclusive.

After collecting a large amount of data, it is necessary to preprocess the data to build the baseline training dataset. It includes removing noisy, redundant, irrelevant, and potentially toxic data. These operations are very important because it has been proven that poor data quality affects the capacities and performance of models [DR5].

Regarding the volume of data required to train such models, collecting, storing and preparing data are the first challenges to solve. They are drivers of the competition and barriers to entry, but also some issues with Generative AI. As the understanding of the data layer is one of the keys to understand the market of Generative AI, we give an in-depth description of it in the paragraph 3.Data: the fuel of Generative AI and of its controversy.

The generative model layer plays a critical role in shaping the model's capabilities and outputs. It is composed of the model architecture backbone, the training procedures, and the model artifacts.

Generative models are called foundation models[2] because they are trained on broad data such they can be applied across a wide range of use cases.

The model architecture defines all the building blocks that compose the deep neural network, and their interconnections. It depends on what are the data types used in input, the data types expected in output, and on the expected capacities of the model.

We differentiate unimodal models from multimodal models.

Unimodal models are tailored to handle a particular type of input data, like text or images, and produce corresponding output of the same type. For instance, text models such as GPT-3 (by OpenAI) take a textual input from users and generate a textual response (see Figure 4). On the other hand, visual models like StyleGAN (by Nvidia) take an input image and transform it based on specified styling constraints (see Figure 5).

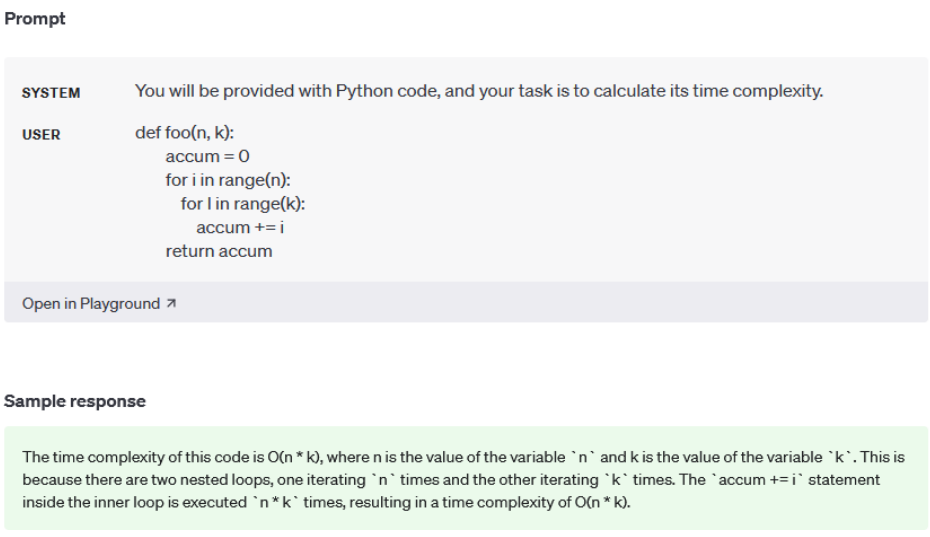


Figure 4 Example of GPT-3 usage to find the time complexity of a program function. Source: Openai playground, https://platform.openai.com/examples

[1] Tokens can be viewed as pieces of words. 1 token represents in average 4 characters in English or ¾ of a word.
[2] Foundation model, a term introduced and popularised by the division of Human-Centered Artificial Intelligence of the Stanford university, 18/08/2021, https://hai.stanford.edu/news/introducing-center-research-foundation-models-crfm

Figure 5 Examples showing how StyleGAN can mix style of several sources (A and B) to generate new faces. Source "A Style-Based Generator Architecture for Generative Adversarial Networks."

Various model architectures, such as Transformer-based, GAN, VAE, and GDM (Generative Diffusion Model), are commonly employed, all relying on similar macro blocks: the encoder block and the decoder block. These architectures may use both blocks, only the encoder block, or solely the decoder block. What unites these models is their unprecedented complexity compared to discriminative neural networks.

It has been shown that **scaling up their architecture plays an important role in increasing the model** [DR5].

This complexity necessitates intricate training procedures and places high demands on computing resources, including processor power, memory, and storage.

Multimodal models are nowadays essentials in Generative AI. Multimodal models are designed to handle a particular type of input data, like text, and produce corresponding output of a different type of data, like image or video. The aim for the model developer is to learn the multimodal connection and interaction from data [DR3]. These models can be used for: text-to-image generation (or vice versa), text-to-audio generation (or vice versa), text-to-music generation, text-to-knowledge-graph generation (or vice versa), text-to-code generation (or vice versa), text-to-molecule generation. Examples of multimodal models are GPT4 from OpenAI, Gemini from Google, StableDiffusion from Stability.

Multimodal models can also combine these capacities, bringing their complexity to higher levels.

The training procedures are the other important components of the generative model layer. When executed, the model actually learns from the datasets regarding the objectives of the procedure, and associates weight and biases in each layer of the neural network architecture.

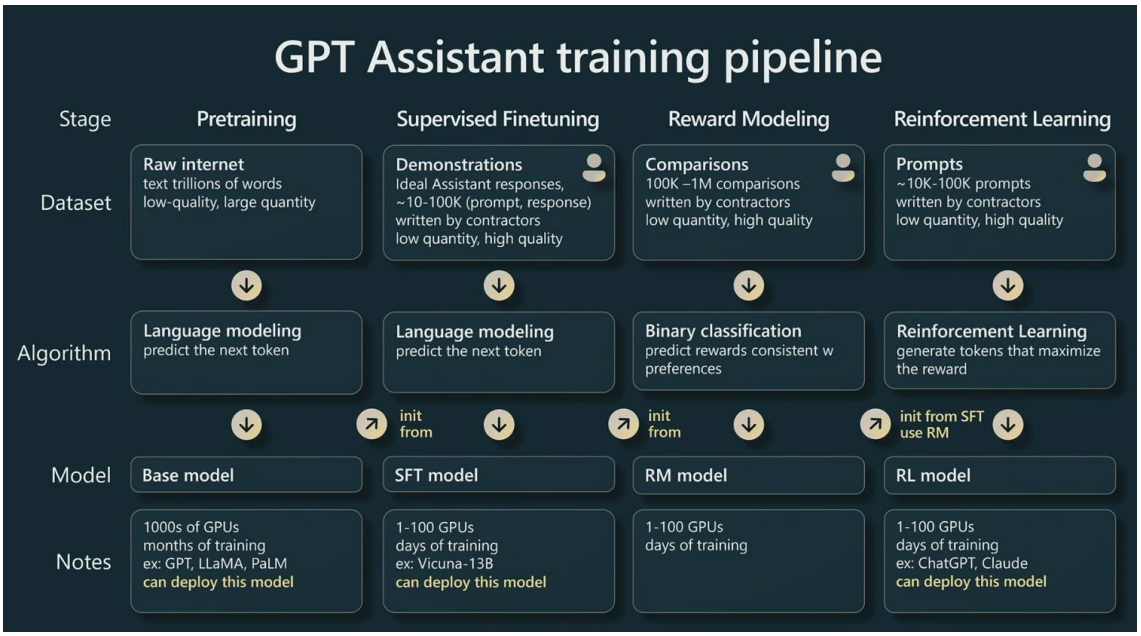


Figure 6 Example of training pipeline for LLMs models serving an GPT assistant (on which user instructions can be issued). Four different training procedures are executed, where each one represents a significative investment regarding talents, computing power and time required. ***Source: The state of GPT, May 2023, Microsoft conference by Andrej Karpathy (OpenAI).***

Because the generative tasks to learn are complex, the training procedures are usually multiples and complex. The model is first trained by using self-supervised technics where only unlabeled data (and so no human-labelled data) are required. During this procedure, huge datasets are leveraged to learn multiple downstream tasks. It usually results in a model with multiple capacities but with average performances.

Then, to increase its performances, the model is usually fine-tuned on a subsample of data labelled with high quality by humans. Finally various reinforcement technics are used to optimize the model following user instructions.

The model artifacts are the products of each step of the training pipeline. Usually, companies only provide the result, the model optimized to follow user instructions, but sometimes also provide the fine-tuned model for further specialization of it on specific domains or use cases by tiers developers.

Training generative models with such complex architectures and large scale requires very talented data scientists and developers to train, benchmark and distribute processing over the infrastructure. This is a considerable barrier to entry on building a Generative model.

The infrastructure layer is essential to understand because it is a driver of competition and is also a barrier to entry regarding its technical requirements and its costs.

The infrastructure layer includes all the hardware infrastructure required to execute the multiple runs of experiment to train and evaluate the model. The hardware infrastructure relies on High Performance Computing (HPC) for massive parallel processing with fast interconnections.

If training a specialized discriminative model is usually done on a single HPC server, locally or in the cloud, trainings of Generative models with billions of parameters are massively parallelized on numerous GPUs[3] or other specialized AI accelerators (like the Tensor Processing Unit by Google). GPUs are interconnected with fast link for data distribution and outcomes aggregation during the parallelized processing. The GPUs are usually aggregated on a server node controlled by a Central Processing Unit. To train a generative model, deep learning researchers and practitioners leverage in the cloud large clusters with GPUs or other AI accelerators (see Figure 7 for an example of cluster). For example, Chinese researchers trained their LLM (named GLM) with 130 billion parameters on a

[3] GPU: Graphics Processing Unit, are specific processors for deep learning model training and inference because they are faster than traditional CPU (Central Processing Unit). They were initially designed to accelerate the 3D rendering of simulations or games on computers or servers. But as 3D rendering and deep learning rely on an intensive usage of linear algebra, GPUs are also able to accelerate computation of deep learning models. In addition, GPUs embed fast and large specific memory units to quickly transfer the data inside the processor. Nowadays, GPUs like the Nvidia H100, are specially designed for deep learning computing only, to increase again their processing speed.

cluster with 768 GPUs Nvidia A100-40G (a high-end GPU with a unit price above 12 K€) during 60 days; or Meta trained its first version of LLAMA by using a cluster of 2048 GPUs Nvidia A100-80G during 21 days [DR5].

The training process of a model consists in optimizing billions of the model parameters by iterating on hundreds or thousands of epochs. During one epoch, the training procedure performs millions on optimization increments by ingesting each time few data samples. During one epoch, the full dataset is ingested. "In consequence, during the model training, huge datasets are ingested entirely several dozens of times (for the recent models like GPT-3, and more for older models). As it would be far too expansive to host a multiple Terra-bytes dataset into a server central memory and into the GPU memory, datasets are saved on disk storages. During the training procedure, the data are transferred from the storage to the central memory by the CPU, and then in smaller batches into the GPUs memory. In consequence, the training infrastructure also requires fast, large storage with distributed file system HDFS. To execute these data transfers and preparation, the infrastructure also requires efficient data pipelines.

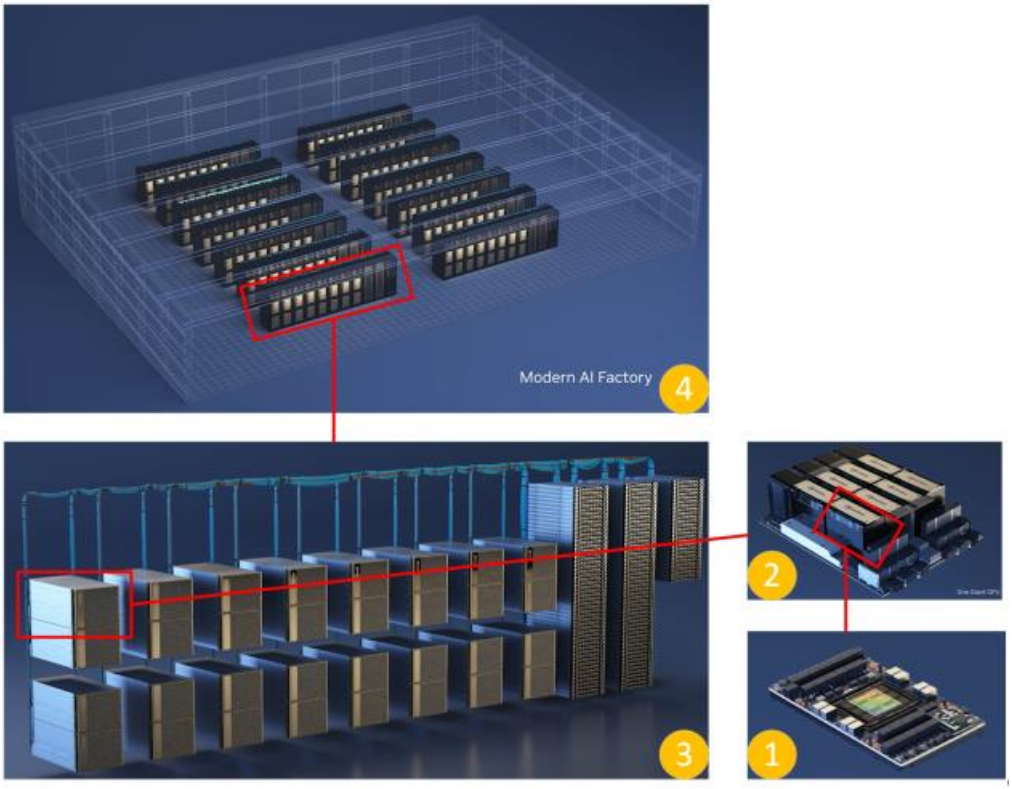


Figure 7 Example of a large cluster of GPUs that can be leverage in the cloud, that Nvidia qualifies as AI factory. This is an example of the high-end Nvidia GPU H100 (illustrated on the subfigure 1), stacked on a server node DGX (illustrated on the subfigure 2) composed of 8 H100 GPU with 640 GB of embedded memory, two intel Xeon CPU, 2 TB of RAM and 30 TB of storage. CPUs execute the operating system and program codes, transfer the data from the storage to the RAM and then to the GPUs via fast specific links, aggregate results and store them back into the storage. DGX nodes are aggregated into groups (called POD) interconnected with fast network connections (see subfigure 3), subfigure 3 shows a cluster of 256 GPUs. PODs are stacked to compose the cluster (see subfigure 4).

The infrastructure layer comprises a software layer, which includes the operating system, GPGPU frameworks (General-Purpose Computing on Graphics Processing Units), deep learning frameworks, and frameworks for managing training experiments, model management, and orchestrating distributed jobs.

GPGPU frameworks, such as Nvidia's CUDA or AMD's ROCm, play a crucial role in accelerating computations on GPUs. These frameworks efficiently distribute and execute computations across the numerous parallelized computational cores within a GPU.

Deep learning frameworks like TensorFlow, PyTorch, and MXNet are foundational components of the software layer. They serve a critical role by abstracting away the intricacies of low-level tasks involved in building and training models. This abstraction facilitates faster, more reliable, and efficient model development.

Due to the increasing complexity of model architectures and training procedures, other specialized software libraries raised as new standard of the industry, like the library Transformers by Hugging Face, to abstract more again operations of model training, usage of pre-trained models for fine-tuning models (specializing a model to a domain or to some tasks). These new specialized libraries tend to lower the barrier to entry for building and experimenting with generative AI models, democratizing access to this technology.

2.3.3 Core Architecture for the model provision

Model developers usually provide access to their models for two kinds of applications:

- Out-of-the-box application: the application and the generative AI model are packaged and do not require any customization. An example of it is ChatGPT, a chatbot interface that uses the GPTs models of OpenAI.
- Custom application build: the application or the generative AI model require some customization.

Depending on the integration capacities they want to offer, model developers provide models as out-of-the-box or customisable.

Models can be deployed to the provider's environment, a public cloud, or a private environment.

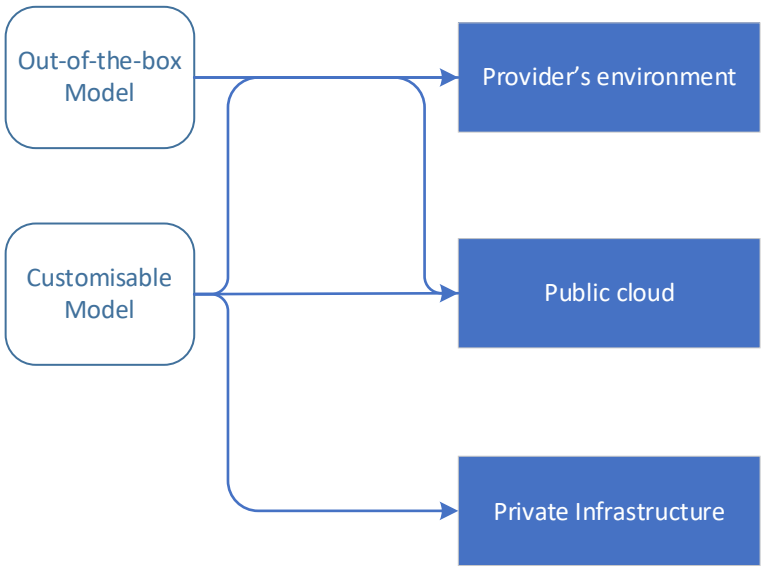


Figure 8 Environments where Generative AI models are deployed.

When models (customizable or not) are deployed in the provider's environment, they are deployed as Model-as-a-Service (MaaS). Users access them through web APIs, the complete architecture to provide the model is hidden to users, enabling easier usage and lower costs. Many model providers like OpenAI, Google, or Stability.ai follow (but not exclusively) this deployment configuration.

Models (customizable or not) are also very often distributed and deployed to public clouds like Azure of Microsoft, AWS of Amazon, G-Cloud of Google, IBM cloud. These cloud providers also offer access to out-of-the-box models of their partners as MaaS. When models can be fine-tuned, these cloud providers give high-level programing interfaces to hide the underlying complexity of the architecture deployed, but the costs to host the customized model remain on the charge of the user. For many customers, these hosting costs may remain prohibitive.

Specific AI platforms like Fujitsu Kozuchi also abstract the cloud provider, and usually integrates additional value like features for fairness assessment, hallucination and attack protections.

Some model providers allow deployment to private infrastructure like private data centers to enable a higher control on security and data privacy. But in this case, in addition to support expansive hosting costs, users have to build the underlying complex architecture to serve the models. Alternately, they can invest in additional tools to build and manage this architecture inside their data center, like Ray by Anyscale or Watsonx by IBM.

Independently of the deployment configuration, the scale of Generative AI models remains a requirement for performances but a challenge for their provision. In any cases, large and scalable clusters of GPUs are leveraged to serve the models.

The components of the core architecture remains the same in any of these configurations, we give a description of them on Figure 1.

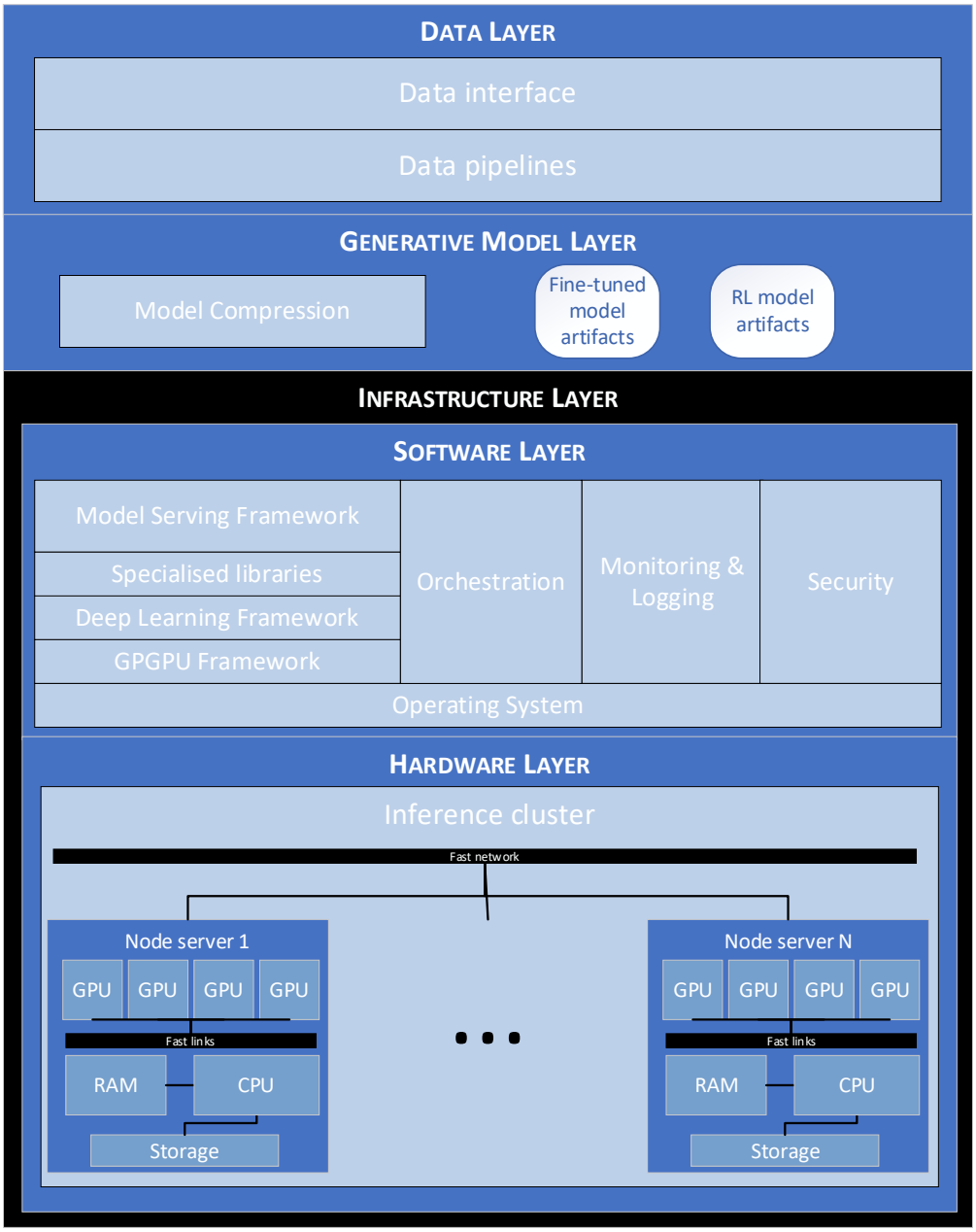


Figure 9 Core architecture for the model provision.

As for the model training, the core architecture for the model provision is composed of three layers: the data layer, the software layer, and the infrastructure layer.

It differs from the model development architecture mostly in the following points:

- The data interface: the objective of the data interface for the model provisioning is not anymore to collect and ingest large and monolithic datasets, but to ingest requests and data from numerous users through various data channels. The data interface usually integrates an API gateway that manages incoming requests, routing them to the appropriate model instance based on routing rules and versioning. It also handles authentication, authorization, and rate limiting. Examples include Amazon or Azure API gateway, or Kong API.
- The model serving framework: replace the model training framework. Its acts as the interface between the model and external users. These frameworks handle model loading (from storage to RAM and to GPU embedded memory), model versioning (to enable the rollout of different versions in production), and efficient request-response handling. The model serving framework needs to manage the complexity of chaining or aggregating multiple model outcomes to produce the outcome to the end-user. Examples include Tensorflow serving, Pytorch serving, Triton Inference server, or Ray serving.
- The monitoring and logging: It monitors the liveness of the service and track the model performance regarding the live data ingested in production. Any deviation on data distribution, or performance metrics are detected to alert of issues. The end-user feedbacks are usually logged. Monitored data are then used to improve the model performance by retraining it.

- The hardware layer: if the inference of the model requires much less computing power and GPUs than the training of the model, the execution of a model instance still requires usually multiple GPUs, especially LLMs or multimodal models. The scale up of the Inference cluster then depends on volume of users, volumes of requests to manage, volumes of data to ingest, and the latency targeted for the content generated to the end-user. Here again, the infrastructure leverages a cluster of GPU nodes, but differs in the sense that the scale varies dynamically with the demand.

Very few initiatives explore the deployment of Generative AI models on lower-end devices like smartphones or personal computers. As examples we can cite the framework Llama.cpp or the research of Apple [DR9]. These initiatives are very new, requires deploying a smaller version of the original model (like for example the version with 7 billion parameters of LLAMA2 instead of the version with 70 billion parameters) with degraded performances. Despite degraded performances, a part of the knowledge of the big model version is distilled in the small model version. Distillation techniques are promising and might enable in the future the deployment of generative AI models on lower-end devices. Now, a few model developers provide smaller versions of their models, and highly technical skills are required to improve the model compression and inference speed. None of these initiatives are yet integrated into a public-grade application.

2.3.4 Core Architecture for model integration into an application

As explained previously, models can be integrated into out-of-the-box or custom applications.

Out-of-the-box applications concern application dedicated to the usage of the Generative AI models, and the AI generated content is the main feature of interest of the application. Some model providers also developed dedicated applications. It is the case of chat applications like ChatGPT by OpenAI, Gemini chat by Google, or again Copilot chat by Microsoft[4]. But any other company than the model provider can develop an application dedicated for AI generated content. For examples, companies like StableDiffusionWeb or StableDiffusionXL created web applications dedicated to image generation from text prompt, using the model Stable Diffusion provided by Stability.ai.

Out-of-the-box applications also concern more usual applications that integrate the content generation as an additional feature. Examples of these applications are Microsoft Office 365 with the integration of their copilot, Adobe Photoshop with new features to edit images and add AI generated content, or GitHub with the integration of their copilot in their code versioning platform.

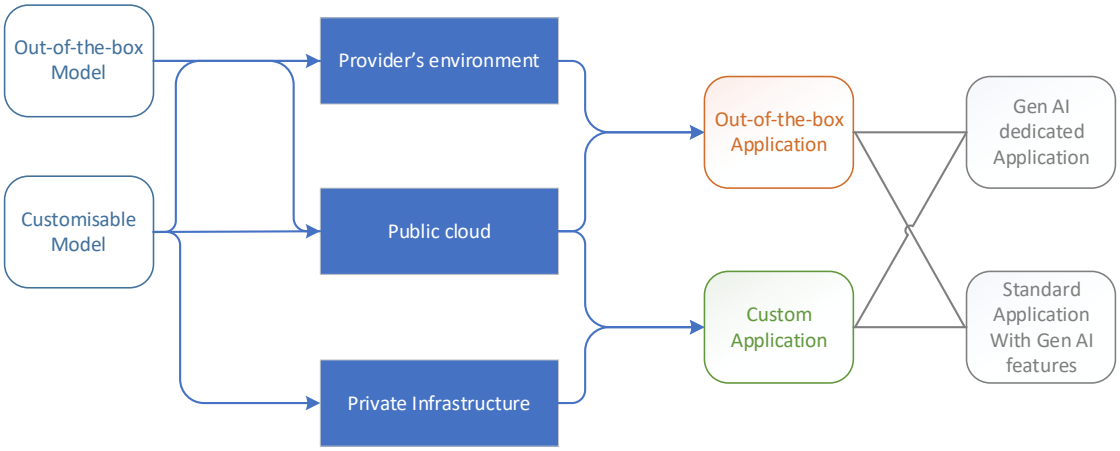


Figure 10 The different types of applications using Generative AI models.

[4] Microsoft is the principal investor of OpenAI and as such has specific conditions on the provision of OpenAI's models.

According to a survey of Altman Solon [DR12], a big majority (70%) of application developers intend to integrate generative AI by using out-of-the-box model. 24% intend to use internally refined models, by fine-tuning models provided off-the-shelf models.

In total 94% of application developers interviewed plan to integrate Generative AI based on off-the-shelf models.

It is to compare with AI models in general, to the 70% of respondents of the survey made by Mc Kinsey in 2020, who declared to have the standard tool frameworks and development processes in place for developing AI models.

It seems that the costs and complexity for developing and hosting Generative AI models is a barrier for the moment.

Development →

Deployment	Application	Out-of-the-Box Application	Custom Application Build	Custom Application Build	Custom Application Build	
	Model	Out-of-the-Box Model	Out-of-the-Box Model	Internally Refined & Nearly Off the Shelf Underlying Model	Internally Built Model	
↑	Application Provider's Environment *(Public Infrastructure)*	**20%**	10%	2%	1%	**32%**
	Public Cloud *(not provider's; Includes Virtual Private Cloud)*	**14%**	10%	2%	3%	**48%**
↓	Private Infrastructure *(On-Premise or Private Data Center)*	1%	**8%**	**8%**	3%	**20%**
		35%	**35%**	**24%**	**6%**	

The core components required to integrate Generative AI models into applications are described in the core architecture diagram Figure 1.

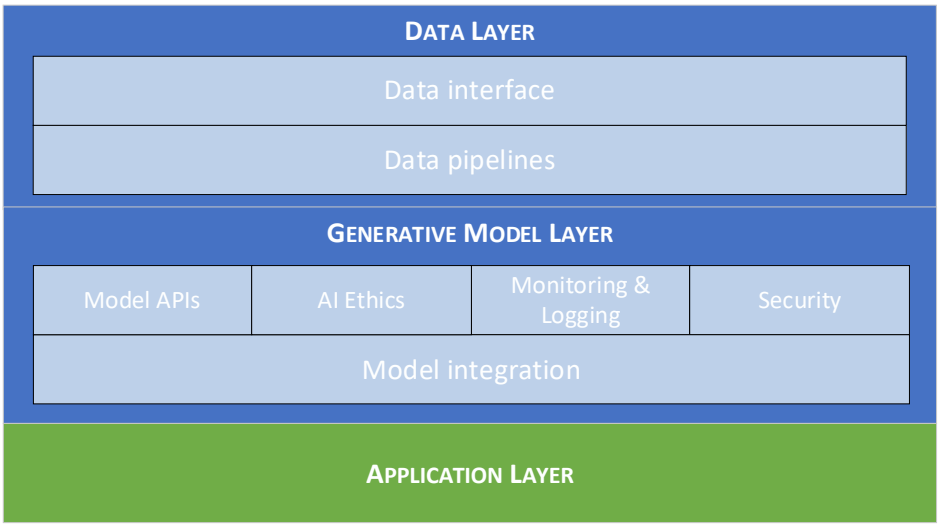


Figure 11 Core architecture for generative AI models integration.

This diagram shows:

- How much the architecture to integrate models is simpler than the one required for model development or model provision.
- The integration of generative AI models still requires data interfaces to collect data inside the application or outside the application via data channels, data pipelines to clean and prepare the data sent to the models.
- The models integration relies on:
  - Calls to at least one generative model via APIs, several models might be needed to fulfil different tasks, but also technical issues inherent to models like LLMs may require multiplying calls to models for the same tasks.

For example, to create chains of thoughts to solve the reasoning issue of LLMs, or to use ensemble of models (eventually from different providers) to solve some hallucination issues or consistency issues. to scall several models.

- AI Ethic tools are useful to grow trust and adoption of users to mitigate fairness issues or hallucination issues of LLMs. Even if model providers integrate some ethic guardians in their models, some issues remain and need to be mitigated by the application developer. This requires knowledge and expertise in AI that the application developer may not have. In consequence, the application developer would have to gather these AI Ethic tools from specific AI platforms like Fujitsu Kozuchi or IBM Watsonx.
- Monitoring and logging is still required to make sure of liveness of services providing models, but also to monitor deviations in data distributions than might become an issue for the model.
- Security is still required for secured data access and management, it is specially important when requests to models reach out to external systems.

2.4 Overview of the technology stack and active players in the market

With the description of the architecture in paragraph 2.3, we have seen that:

- Many components are required in the technology stack to develop, provide, or integrate generative AI models.
- Large scale of models became a requirement for performance with many consequences on the technical requirements and costs of generative AI models.

These two statements have major impacts on the market structure.

We give on Figure 12 an overview of the technology stack involved in the architecture of Generative AI for model development, model provision, and model integration into applications. This overview show for each technological layer, the active players in the market. The given list of active players is not exhaustive and gives examples of the most renown players.

Some players like Google are active at different layers of this stack even if they don't appear systematically in the different layers of this figure. Following the figure, we give some comments on each layer and active players.

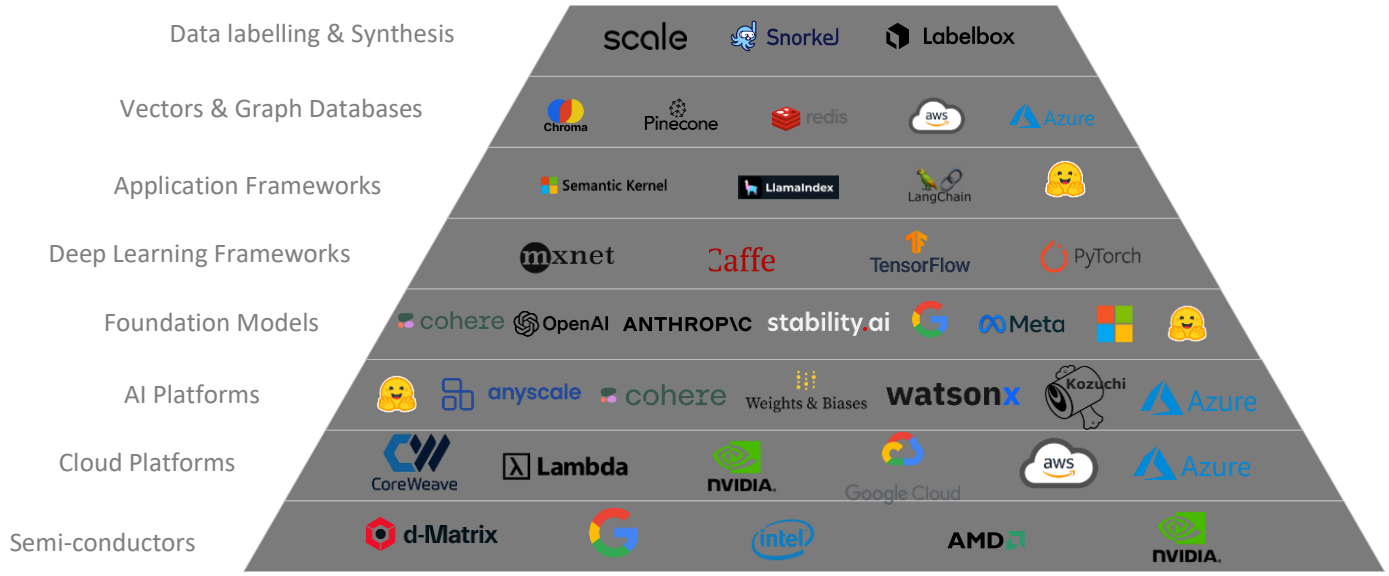


Figure 12 Overview of the technology stack and active players.

- Semi-conductors: essential parts of the hardware infrastructure, they execute applications and model computations. They include processors from manufacturers like Intel or AMD, and AI accelerators (GPU or TPU) by manufacturers like NVidia, AMD, Intel, Google. The very high and increasing demand of AI accelerators enable the emergence of new players like D-Matrix that focuses on balancing computing acceleration and energy consumption. Google anticipated a long time ago the increasing demand on AI accelerators and designed its own TPUs for internal use and commercial use within its cloud platform. Nvidia dominates 90% of the selling

market of AI accelerators. The AI demand and the dependency to Nvidia is so high that actors like Microsoft and OpenAI think to design their own AI accelerators.

- Cloud platforms: they are the platforms used to leverage the large clusters of GPUs (or TPUs) required to train and infer models. They deploy servers embedding the semi-conductors. Big cloud actors like Microsoft Azure, Amazon AWS, or Google cloud are general-purpose cloud providers but can deploy the complex infrastructure for foundation models. General-purpose cloud providers provide all the tools of above layers. Nvidia is a strategic partner of most of cloud providers. The high demand on AI enables the emergence of new players like CoreWeave or Lambda specialized in the deployment of large GPU clusters.
- AI platforms: platforms to train, deploy, infer, monitor, secure models. Most of the general-purpose cloud providers like Microsoft Azure, Amazon AWS and Google cloud provide their own AI platforms. They also partner and deploy specialized AI platforms like IBM Watsonx or Fujitsu Kozuchi or Weight & Biases. These ones add value for AI developers by abstracting the complexity for model development and monitoring, or AI ethic management. The new player Hugging Faces is unique, it proposes access to an open-source AI community, with the distribution of thousands of models created by (very) small actors or (very) big actors like Meta.
- Foundation models: include all the generative AI models. This is probably the layer where the biggest competition is ragging. The new AI actor OpenAI, largely funded by Microsoft, leads the charts with its GPT-series, DALLE, or SORA. Big players like Microsoft, AWS, Google, Meta develop and provide their own foundation models for internal use or commercialisation. New players like Anthropic, Cohere or Mistral AI emerge at a high frequency. The market is very dynamic and shared between closed-source and open-source actors.
- Deep learning frameworks: the basic frameworks to train and deploy deep learning models, include generative AI models. Big actors like Tensorflow (by Google), Pytorch (by Meta), Caffe or MxNet (by Apache) are in place since many years. Tensorflow and Pytorch dominate the market.
- Application frameworks: are the specialised libraries used to build applications integrating generative AI models. They aim to abstract the complexity or load complex models or chaining them. HuggingFaces built a widely used library to use models based on the transformer architecture. Actors like Langchain, LlamaIndex or Semantic kernel (by Microsoft) emerged with the success of LLMs and aims to simplify their connection to data sources, and models inter-connections.
- Vector & Graph databases: are specific databases to build complex and performant knowledge databases. New actors like Chroma, Pinecone or Redis emerged, but big actors like Microsoft Azure and Amazon AWS also developed their specific products. LLMs make use of all the power of these knowledge databases.
- Data labelling & synthesis: foundation models are very big data consumers. Some of them are supervised during their training and require labelled data. All of them require labelled data for fine-tuning. Scale.ai dominates the market and made it success by labelling the data used for autonomous driving models. Scale.ai is now very active in the generative AI market. But even in our digital area, available real data aren't enough to feed the foundation models. There is an emerging market to create synthetic data to complete the model training.

This diagram reflects the complexity of Generative AI by showing the many technological layers of its composition. Its pyramidal shape also shows the greater importance of some layers on which relies other layers. Semi-conductors and cloud platforms are the foundation layers of this technology stack required to train and provision the generative AI models. The big cloud providers like Microsoft, Amazon and Google have the financial, human, and technical resources to develop and provision their own models.

Few actors have the required resources to develop and provision their own models. There exists a high risk of model lock-in for application developers. This risk is now mitigated for LLMs in the language domain thanks to framework like Transformers of HuggingFaces or Langchain that enable to switch more easily from one model to another one. But as each model has its own characteristics and performances, the risk of model lock-in is not eliminated.

The big cloud providers also developed their AI platforms that integrates all necessary tools and resources of the above layers in the pyramid. For model and application developers there exists a high risk of platform lock-in. To mitigate this risk developers must invest in the integration of specific tools to become platform agnostic.

It is interesting to note that the attractive market of Generative AI push big tech companies like Meta to deviate from their core strategy. Meta is a very active player of Artificial Intelligence since already many years, but until recently had for strategy to use its AI internally to analyze the large data volume flooding its social networks and serve advertising publisher with AI outcomes. Meta was used to publish the code and the artifacts of its models, but artifacts could not be used for commercial purpose. With the raise of generative AI Meta changed its strategy and now distribute some of its models for commercial use.

3. Data: the fuel of Generative AI and of its controversy

Data plays a pivotal role in generative AI, acting as the foundation for model training and influencing both its capabilities and potential limitations. Here's a breakdown of its significance and key characteristics:

Role of Data:

- Training Source: Generative AI models "learn" by analysing large datasets of text, images, code, or other relevant information. This data informs the model's understanding of patterns, relationships, and stylistic elements, allowing it to create new, original content.
- Performance Influence: The quality and quantity of data directly impact the performance of generative AI models [DR5], [DR15]. More data often leads to higher accuracy, diversity, and coherence in generated outputs.
- Bias Reflection: Data can harbour biases based on its source or collection method. If these biases are not addressed, the generative model may perpetuate them in its outputs, raising ethical concerns.

Relevant Data Characteristics:

- Volume: Larger datasets lead to better results, but diminishing returns occur beyond a certain point.
- Diversity: Varied and representative data ensures the model can generalize to unseen examples and avoids biases towards specific styles or patterns.
- Quality: Clean, accurate, and error-free data is crucial for reliable model training and output generation.
- Relevance: Data needs to be relevant to the specific task or type of content the model is generating.
- Accessibility: Access to relevant and high-quality data is a significant challenge for some developers, especially regarding proprietary datasets.

A review of scientific publications shows that common public datasets are widely used to train generative AI models. These public datasets are composed by collecting public data of diverse kinds by scraping the web or compiling books or articles (see Figure 13). When the model developers complete their data from other sources, they are not always clear about their methods and sources. The use of training datasets for generative AI has sparked several controversies about web scraping and use of copyrighted content, or biases reinforcement.

A few legal actions have been filed against generative AI companies including Stability AI in the High Court of Justice in London, in California and against GitHub, Microsoft and Open AI. If successful, these legal actions could hamper the progress and development of generative AI, including restricting the use of copyrighted content as training data [DR17]. The author of [DR17] advice to explore new laws for AI to answer concerns like:

- Generative AI are taking advantage of the fruits of human labour? Have we upset the apple cart?
- We have the right sense of fair-play, recompense, or attribution for the use of content.
- The rationales and assumptions on copyright have been undermined including on concepts of ownership, attribution and creativity?

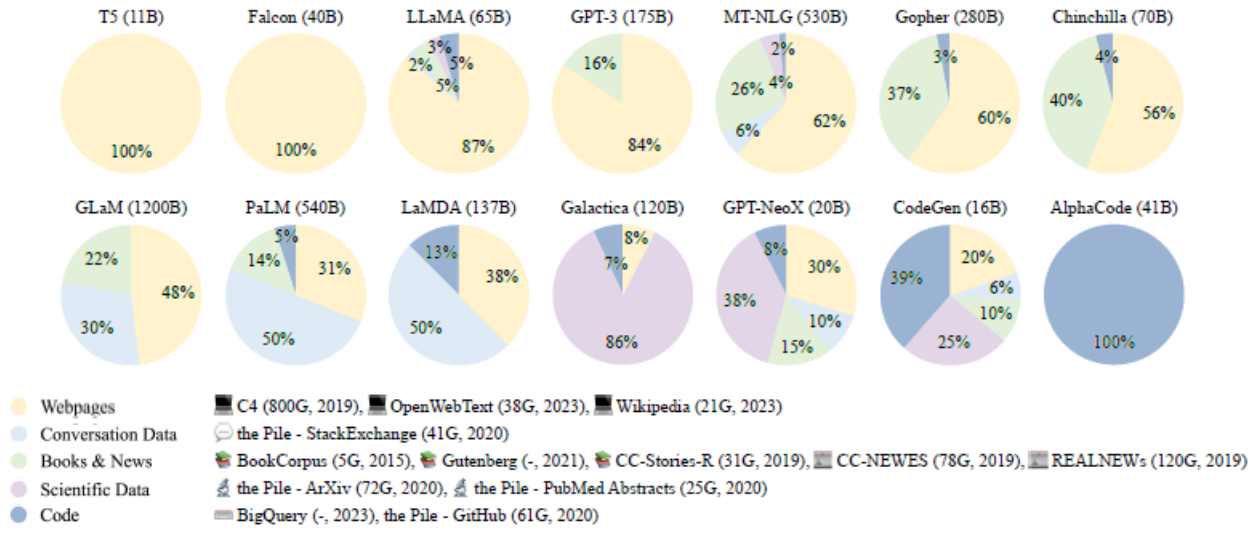


Figure 13 Data sources and their ratio in the pre-training data for existing LLMs [DR5].

Biases of information are reinforced by generative AI because biases contained in datasets are learned by models and models reflect them into their predictions. Algorithms of models can even amplify biases and show themselves biased behavior due to certain design choices. The outcomes of biased algorithms can be fed into applications and affect users' decisions. As users feedbacks are used to fix and optimise models, biases induced by users are then injected again in the loop [DR18].

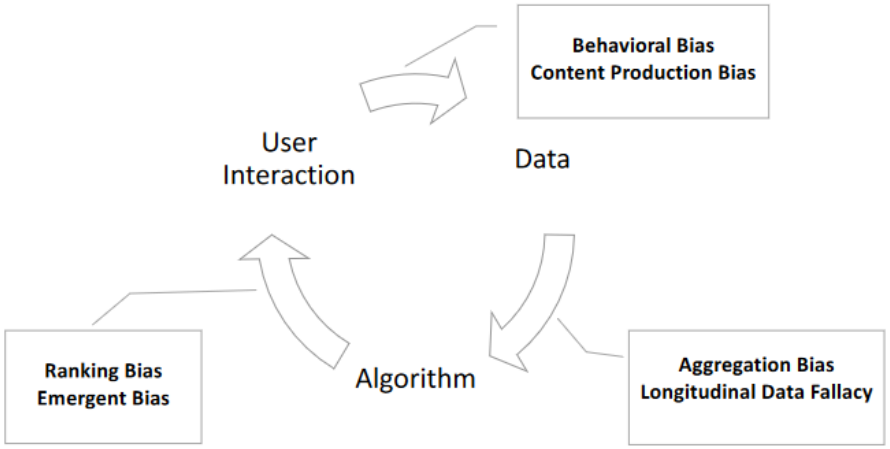


Figure 14 The loop mechanism of bias reinforcement.

If model developers are aware of biases in machine learning and of mechanisms of bias reinforcement, scales of datasets and models make of the model fairness a real challenge and difficulty. Despite their expertise and means, even the big techs like Google or Microsoft have issues to handle fairness and biases before the product releases[5].

But as generative AI may create and spread content everywhere on the web, in enterprises, the new risk is too spread biases and misinformation at unprecedent scales and rhythms.

[5] Gemini image generation got it wrong, https://blog.google/products/gemini/gemini-image-generation-issue/ Buzzy ChatGPT chatbot is so error-prone that its maker just publicly promised to fix the tech's 'glaring and subtle biases', https://fortune.com/2023/02/16/chatgpt-openai-bias-inaccuracies-bad-behavior-microsoft/

4. Market Analysis & Competition

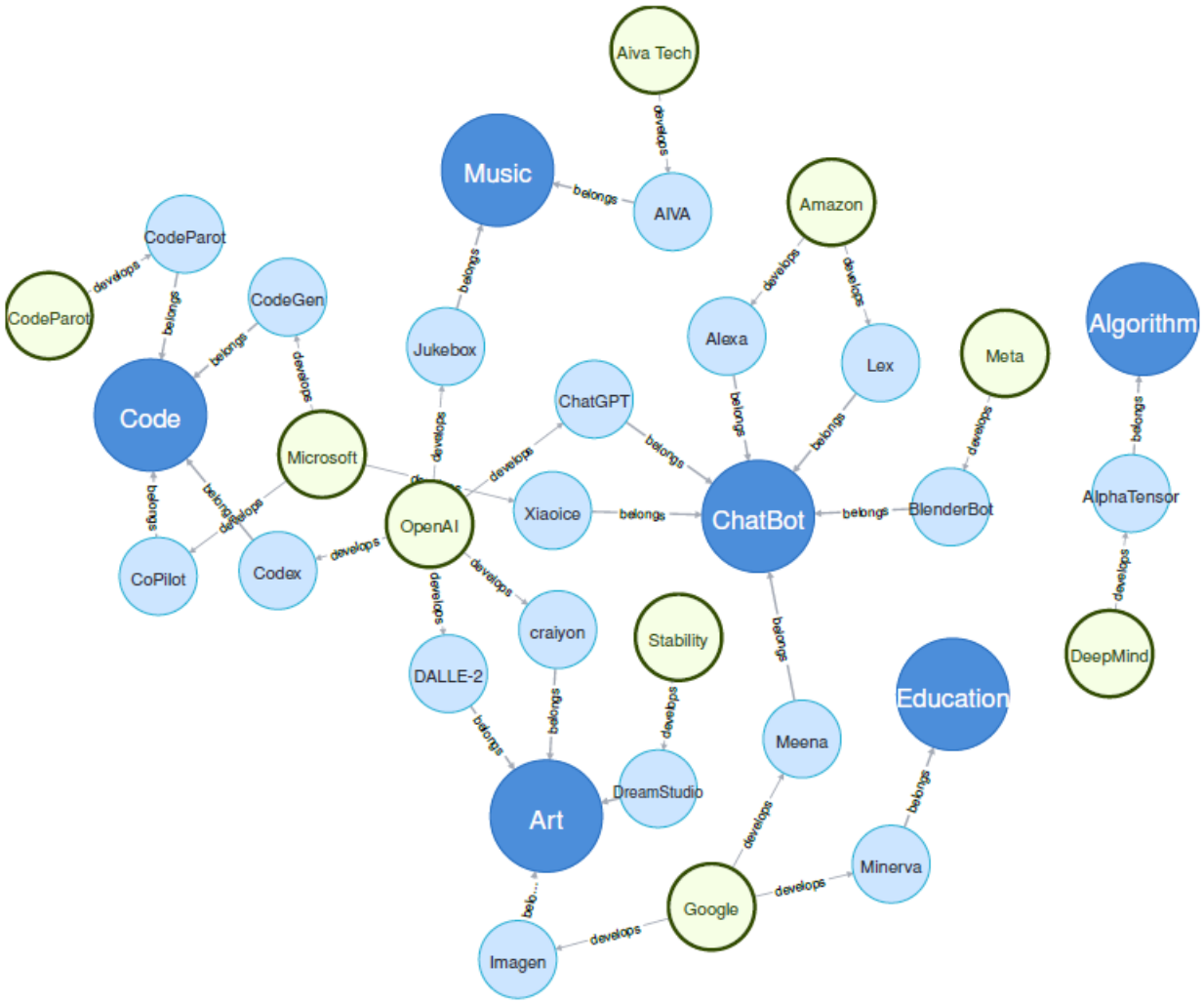


Figure 15 A relation graph of a current research areas, applications and related companies [DR3], where dark blue circles represent research areas, light blue circles represent applications and green circles represents companies.

Despite it is still early days, the market of Generative is very dynamic. Gartner estimated the market revenue to 24 billion of dollars in 2023, and projects an increase to 60 billion of dollars by 2026. A survey of McKinsey estimates that generative AI could add up to $4.4 trillion annually on the global economy.

Generative AI will have a significant impact across all industry sectors, and the potential to change the anatomy of work.

The complex architecture of generative AI models, combined with the immense requirements for data, computing power, and financial resources, makes it a challenging market to enter. This complexity gives established technology companies a significant advantage, as they can leverage their vast financial and technical resources to develop their own generative AI technologies and form strategic partnerships. Additionally, large cloud companies have a presence across the entire technology stack, potentially capturing a significant portion of the market revenue.

We propose hereafter a market and competition analysis of Generative AI.

4.1 Drivers of Competition

The analysis of the core architecture of Generative AI systems and of the role of data highlighted the following main drivers of the competition.

- **Datasets and data quality**: Access to diverse, high-quality data is crucial for training effective Generative models. Companies with access to vast internal or proprietary datasets often have an edge (e.g., Google, Facebook). Leveraging public datasets remain necessary to develop new models, as they serve as benchmarks to understand and measure performance evolutions. But it also a risk as public datasets are not free of biases and

copyrights. To mitigate these risks, companies need to collect data elsewhere and constitute new private datasets. Regarding the very large size (for example, OpenAI trained GPT3 on 45 Terra Bytes of data[6]) of datasets required, it requires tremendous efforts and skills (to ensure data quality and minimization of embedded biases).

- **Talents:** With the increasing complexity and scale of models, more than before highly knowledgeable and skilled talents are required to train, deploy. **The stake is particularly on data-scientists, machine learning engineers and developers training the new models**. As mentioned by the authors of [DR1], the salaries for these profiles usually go into six-digit territory that only some of the most well-funded companies are able to pay. Regarding the integration of generative models into applications, as models produce outcomes in a form readable by non-AI-experts, there is a shift from data-scientists and machine learning engineers to software developers, as observed in Mc Kinsey survey in 2023 [DR19]. For use cases where the usage of multiple models is required, data-scientists and machine learning engineers might still be required to integrate models into applications. Globally, hiring for AI-related roles remains a challenge (see Figure 16). Machine learning engineers and data scientists, the roles that build AI models, are still the most difficult profiles to hire.
- **Generative AI models**: are the engine of applications and services using generative AI and give unprecedent capacities to non-AI-experts users. The sudden success of applications like Stability XL for image generation or ChatGPT for text generation, pushed generative AI models to the forefront**. Despite it is still early days, a vast majority of people already tried them at least once, and a big proportion of people use them regularly at work or outside work** (see Figure 17). It led to a raging competition to develop and provide models (see Figure 2).
- **Access to AI accelerators**: is essential to train or customize generative AI models. Companies who want to enter the race of Foundation models development, need to leverage large clusters of GPUs or TPUs both for training and provisioning models. **A survey made by CBInsights shows that 80% of the 14.1 billion dollars funded in generative AI for 2023 were attributed to build new generative AI infrastructure [DR26]**. The demand on AI accelerators became so high that cloud computing providers are facing challenges meeting the demand [DR20], the ongoing shortage reflects the sudden demand for ultra high-end GPUs [DR21].
- **Model accessibility**: is part of the competition around model provision. For model providers it is important to create well designed and documented APIs with attractive pricing models.
- **Ease of integration and value creation**: offering pre-trained models and templates for specific use cases lower the barrier to entry for application developers and accelerates adoption. Allowing (expert) users to fine-tune models and non-expert users to control the content generation creates added value and foster differentiation from other providers.
- **Adoption of Generative AI for enterprises**: A survey of McKinsey [DR10] estimates that **generative AI could add up to $4.4 trillion annually on the global economy, across 63 use cases**, increasing the impact of AI by 15% to 40%. It states that generative AI will have a significant impact across all industry sectors and that it can significantly increase labour productivity. Banking, high tech and life sciences are the industries that could see the biggest impact on their revenues.
- **Adoption of Generative AI for individuals**: generative AI can support individuals in creating content of quality at work and outside. According to [DR10], Generative AI has the capacity to reshape the nature of work by enhancing individual workers' abilities through the automation of certain tasks. Current generative AI and related technologies can potentially automate 60 to 70 percent of employees' current work activities, a significant increase from the previous estimate of automating half of the work time. This acceleration in automation potential is attributed to generative AI's improved proficiency in understanding natural language. The biggest impact is on knowledge-based tasks, specifically those related to decision-making and collaboration. Authors of [DR10] see generative AI having the most impact on activities of educators, professionals and creatives.
- **Deployment of foundation models to lower-end devices**: as distillation and compression techniques should continue to improve with time, and the computing power of lower-end devices like smartphones should continue to grow, there will be a moment when deploying foundation models on devices like smartphones will be feasible. We think that it will become a flagship functionality of such devices to provide the next generation of smart assistants that do not require to send requests and data over the internet and risks data leakages. Apple and Qualcomm are actively working on smartphone deployment, Nvidia is actively working on personal computer deployment.

[6] https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/

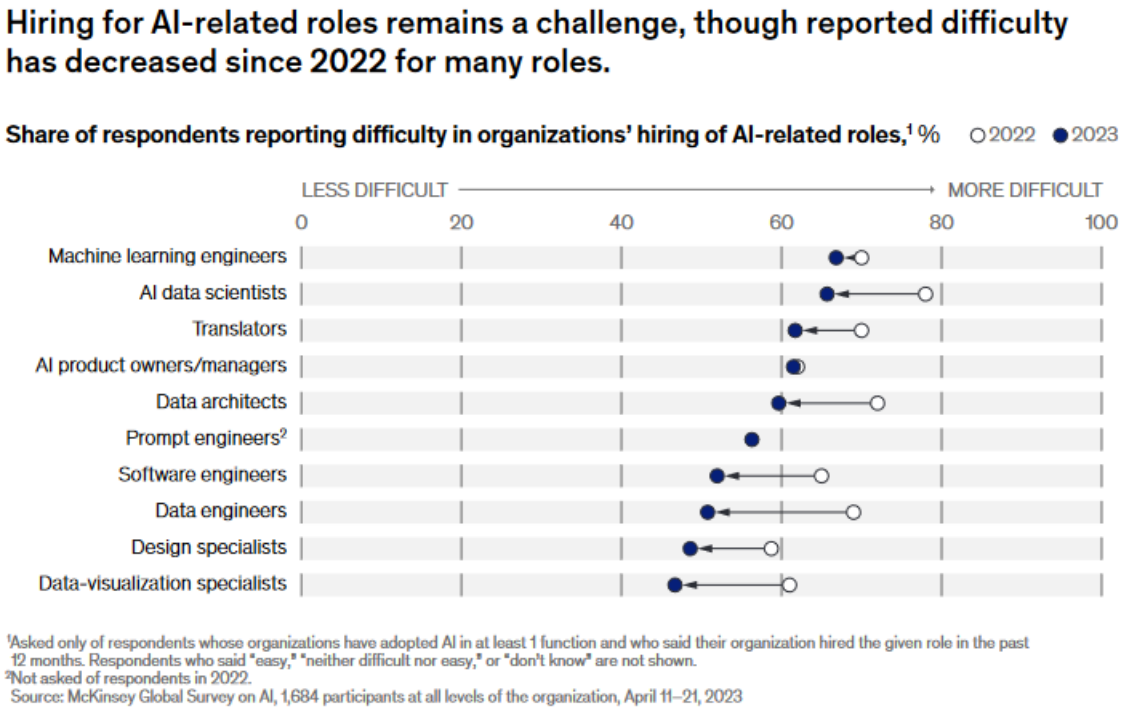


Figure 16 Hiring for AI-related roles remains a challenge [DR19].

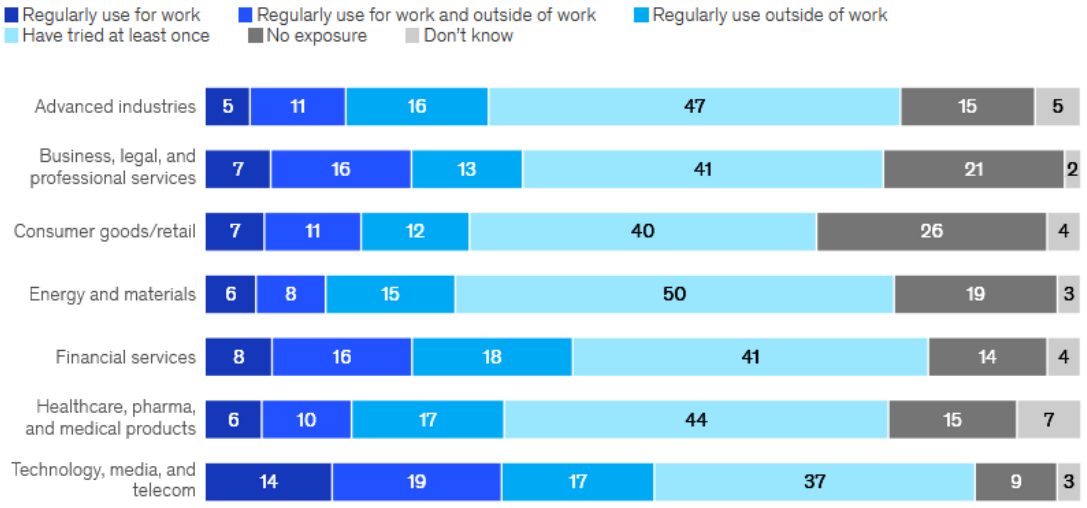


Figure 17 Early days of Generative AI, but its use is already widespread. Source: Mc Kinsey, April 2023

4.2 Barriers to Entry and Expansion

The analysis of the core architecture of Generative AI systems and of the role of data highlighted the following barriers to entry and expansion.

- **Costs of the infrastructure required for training models**: Understanding the expenses associated with infrastructure for training large-scale generative AI models is crucial for assessing the barriers to entry and expansion in this field. However, obtaining specific cost details is often challenging as companies tend to keep such information confidential. For models like GPT-3 or LLAMA2, the costs for a single training run are estimated to be in the millions. Multiple runs are typically necessary to optimize the model for public release. OpenAI's cofounder, Sam Altman, mentioned at an MIT event in July that the cost of training foundation models like GPT-3 exceeds $50 million to $100 million and is increasing over time with new model versions. See Annex B for further explanations on scale of foundation models and their training costs.
- **Costs of the infrastructure required for provisioning models**: here again information is kept confidential by companies. To give some insights on the costs of the hosting infrastructure, we can take the example of the LLM

Falcon with 180 billion parameters. Amazon AWS recommends[7] to use their compute instance ml.p4de.24xlarge with a listed price of almost 33 USD per hour (on-demand). Such a deployment would cost at least 23,000 USD per month (list price, if the application runs every day 24 hours). With a latency of 45 ms per token for this model, this compute instance can generate only 22 tokens (29 words) per second, and 1740 words by minute, a bandwidth for maybe 1 to 10 user requests by minute, or 60 to 600 user requests by hour. This bandwidth is not enough to support hundreds of users by hour, so the deployment configuration would require multiples compute instances of this kind. Taking the example of ChatGPT by OpenAI with probably at least 1 million active users by hour and assuming that each of them only submit one request by hour, this would cost OpenAI more than 38 million dollars to host the model behind ChatGPT in AWS with the ml.p4de.24xlarge compute instance. It is therefore easier to understand that supporting the infrastructure costs for hosting generative AI models, is a barrier to entry or expansion for small or medium companies who like to host models.

- **Costs for fine-tuning models**: customizing models to increase their performances on specific domains may be necessary. Developers who need to customize models have two different options: via a dedicated API to send custom data and request for fine-tuning computation or by loading the model file artifacts in their own infrastructure. The second one again will most of the time be a substantial barrier to entry for many developers. The first one is more affordable for many developers: for example, OpenAI give a specific API to fine-tune its model GPT-3.5-turbo, at a price of $8 for one million tokens. If 100000 prompts are required for fine-tuning, with an average of 250 tokens per prompt, this would lead to a cost of $200, a cost affordable for many companies.
- **Talents**: are a barrier to entry or expansion in the development of generative AI models. It requires highly experienced and talented data-scientists and machine learning engineers, two profiles difficult to hire. In addition, foundation model development requires large teams composed of many data-scientist, machine learning engineers, software developers with salaries that only the most well-funded companies can pay [DR1].
- **Datasets**: are a barrier to entry or expansion in the development of generative AI models. Very large datasets are required to train such models. Some public datasets can be used but most of the time are still biased, may embed copyrighted content, or cannot be used for commercial purposes. In consequence, companies who want to build models must leverage proprietary data and constitute their own datasets with tremendous efforts.

4.3 Emerging Competition Issues

With the analysis previously in this report, we foresee the following potential issues for competition.

- **Access and control of resources**
  - Computing resources and AI accelerators: there exists a high risk of limited access to high-end AI accelerators and of their control by big technological companies (including American and Chinese companies). The on-going shortage on high-end GPUs and the increasing pushed these big companies to secure their GPU provision by buying big stocks or renting them via third parties like CoreWeave [DR22]. Companies like Microsoft, Google, Meta or Amazon are now trying to mitigate the risks of their dependency to Nvidia by building their own AI chips. Amazon and Google have already started to commercialize their own AI chip inside their cloud offering.
  - Datasets: Generative AI relies heavily on datasets as its primary source of information. However, publicly available datasets often have biases or contain copyrighted material, making them challenging for commercial model development. Companies possessing extensive proprietary datasets gain a competitive advantage. A noteworthy trend in the market is the monetization of data access by certain social networks, such as Reddit. Model providers can leverage content generated by platform users to create new private datasets. The growing demand for private data introduces a potential concern: social network platforms may unilaterally assert ownership over user-generated data, claiming exclusive rights to derive new benefits from this data.
- **Dominance of few model providers**: The difficulty of entering the development arena for foundation models is currently very high. As the complexity and scale of these models continue to increase, the entry barrier is expected to rise even further. Consequently, only a limited number of global companies can afford to invest in the research and the development of foundation models. This situation poses a significant risk of reinforcing the dominance of these few companies in the future.
- **Network effects and platform dominance**: Cloud providers provide access to GPU clusters, foundation models, and integrate all necessary hardware and software resources for developing generative AI applications. Leveraging network effects, these providers stand a significant chance of emerging as the predominant winners in the generative AI market.

[7] https://aws.amazon.com/fr/blogs/machine-learning/falcon-180b-foundation-model-from-tii-is-now-available-via-amazon-sagemaker-jumpstart/

- **Lack of transparency in data used and algorithms**: Many model providers in the market operate as closed-source entities, refraining from sharing the program code, model file artifacts, or details regarding the datasets used for model training. Research publications typically offer limited technical insights into the model's architecture, training methodologies, and performance benchmarks on public datasets. However, such information falls short of providing transparency on the performance of commercial products. To transition into commercial products, model developers often employ private datasets, additional training procedures, and extra code to enhance model security and ethics. This lack of transparency poses challenges in auditing closed source and commercial model versions for data privacy, fairness, and security. Users must place trust in providers' commitment to ethical and security policies, although instances have demonstrated unintended generation of inappropriate, biased, or racist content by users, both expert and non-expert. Notably, efforts to mitigate bias before model release have not completely eliminated the issue, prompting the need for policymakers and regulators to consider strategies addressing businesses utilizing language models. The efficacy of providing models with additional context as a solution to bias is also questioned in a study referenced as [DR23]. To address risks related to data privacy, security, and biases, [DR24] proposes a comprehensive auditing approach comprising three layers: governance audits, model audits, and application audits. The authors emphasize that technology providers bear the responsibility for ensuring the legality, ethics, and technical robustness of Language Models (LLMs), presenting both moral and material incentives to undergo independent audits. In the long term, the authors suggest that technology providers establish and fund an independent industry body dedicated to conducting or commissioning governance, model, and application audits. Policymakers can play a role in facilitating the emergence of an institutional ecosystem capable of carrying out and enforcing such audits for LLMs, according to the recommendations in [DR24].

4.4 Open source versus proprietary Generative AI

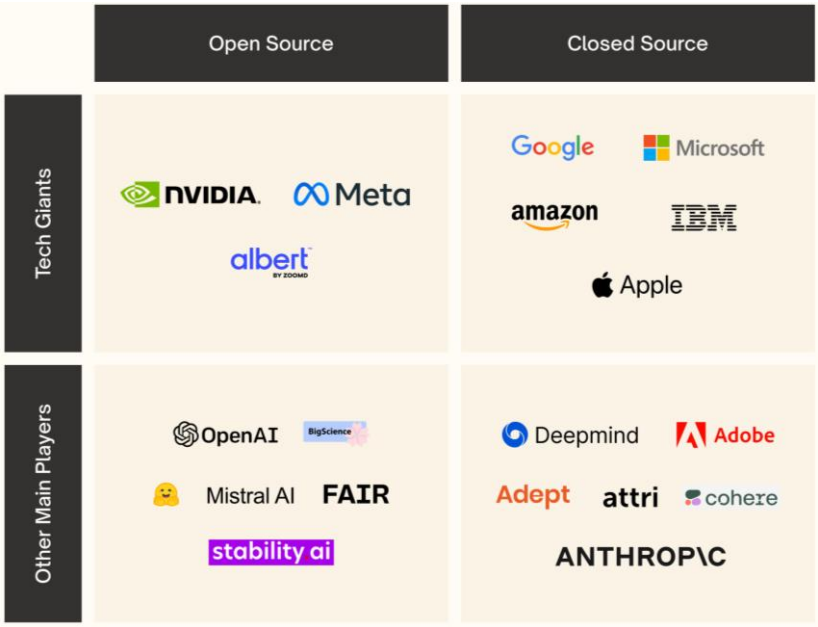
The Generative AI space has seen an exponential rise with the emergence of startups and other players. However, making an informed decision in this nascent field extends beyond mere brand recognition or price comparisons. A critical aspect of this decision-making process involves choosing between open-source and proprietary foundation models.

The market analysis for model provision shows that it is shared between closed-source providers, or open-source providers.

Most the technology giants like Google, Microsoft, Amazon, IBM or Apple are closed-source providers. Exceptions in this perimeter are Nvidia and Meta.

Nvidia do not provide their models with payable licenses, they sell GPUs and have for commercial strategy to ease the development of software with free licensed products, with the only requirement of using Nvidia GPUs. With this strategy, Nvidia aims to increase the developer's dependency to its hardware products.

Meta distributes since many years their AI models as open-source and free licensed software. But until recently their licenses allowed only personal and research usage. With the release of their LLM LLAMA in version 2 they changed their strategy by allowing also commercial usage.



OpenAI started its history as "a non-profit artificial intelligence research company with the goal to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return."[8] Prior to the release of their Language Model (LLM) GPT-3, all OpenAI models were open-sourced, provided without license costs, although users could opt for a paid Model-as-a-Service (MaaS) through payable APIs. The development of the expansive GPT-3 model, requiring substantial funding because of its very large scale, prompted a shift in their provision strategy. Presently, OpenAI's new models are closed source, and users incur license fees through API access to these models. Microsoft stands as the principal investor in OpenAI, having contributed a substantial funding of 13 billion dollars to support its initiatives.

The French company Mistral AI, created in 2023 by former researchers of Google and Meta, developed and distributed very promising models free of license costs, but limited to smaller versions (with less than 15 billion of parameters). Mistral AI has recently released a large version of its model (the number of model parameters is kept confidential) able to compete with models of OpenAI and Google leading the performance board. This model is released with license fees charged through the API access to the models. Mistral AI also signed a pluriannual partnership with Microsoft. Mistral AI will commercialize its large model as MaaS in the Azure platform, and thanks to this partnership will access to the supercomputing infrastructure for AI of Azure, benefits of large scale to market, and collaboration with Microsoft on research and development[9]. By multiplying partnerships with providers, Microsoft may try to mitigate its initial dependency to OpenAI.

OpenAI and Mistral AI serve as examples illustrating that, despite companies initially sharing some models through open source, there is no assurance or commitment to persist in this approach in the future. Additionally, they highlight new strategic moves like did Microsoft, leveraging access to its supercomputer to establish exclusive partnerships with certain model developers.

On our knowledge, **there exists very few initiatives releasing open-source models including the code and the model file artifacts, with license free of charges and that allow a commercial usage.** Examples of such model providers are Meta with LLAMA2, Stability.ai with Stable diffusion and BigScience with Bloom.

These initiatives are important to stimulate the competition on the market because they enable to lower the entry to barrier. Companies that want to invest the market but that cannot afford research and development costs and model training costs, can use the model file artifacts, and invest in the provision infrastructure for internal use or commercial use.

As underlined by the author of [DR25], **open source in AI is more than just access to the source code**. The backbone architecture, data and procedures used to define the weights of the neural network during the model training are also required to open-source the AI model. Without them, the training outcomes cannot be reproduced, and models remain black boxes lacking transparency.

Open sourcing the foundation models can help in making models more transparent and accountable [DR23].

Open source in AI is also important to stimulate the research and innovation. Authors of [DR5] explain that in the case of LLMs, given the huge cost of model pre-training, well-trained model artifacts are critical to the study and development of LLMs for the research community. Open-sourcing AI models seems necessary to continue the research outside of the only companies that can afford pre-training costs. The example of the release of LLAMA by Meta shows how much it stimulated the research in the field of LLMs, being the roots of much research applied to domains of mathematics, medicine, law or finance (see Figure 1).

[8] https://openai.com/blog/introducing-openai
[9] https://encord.com/blog/mistral-large-explained/

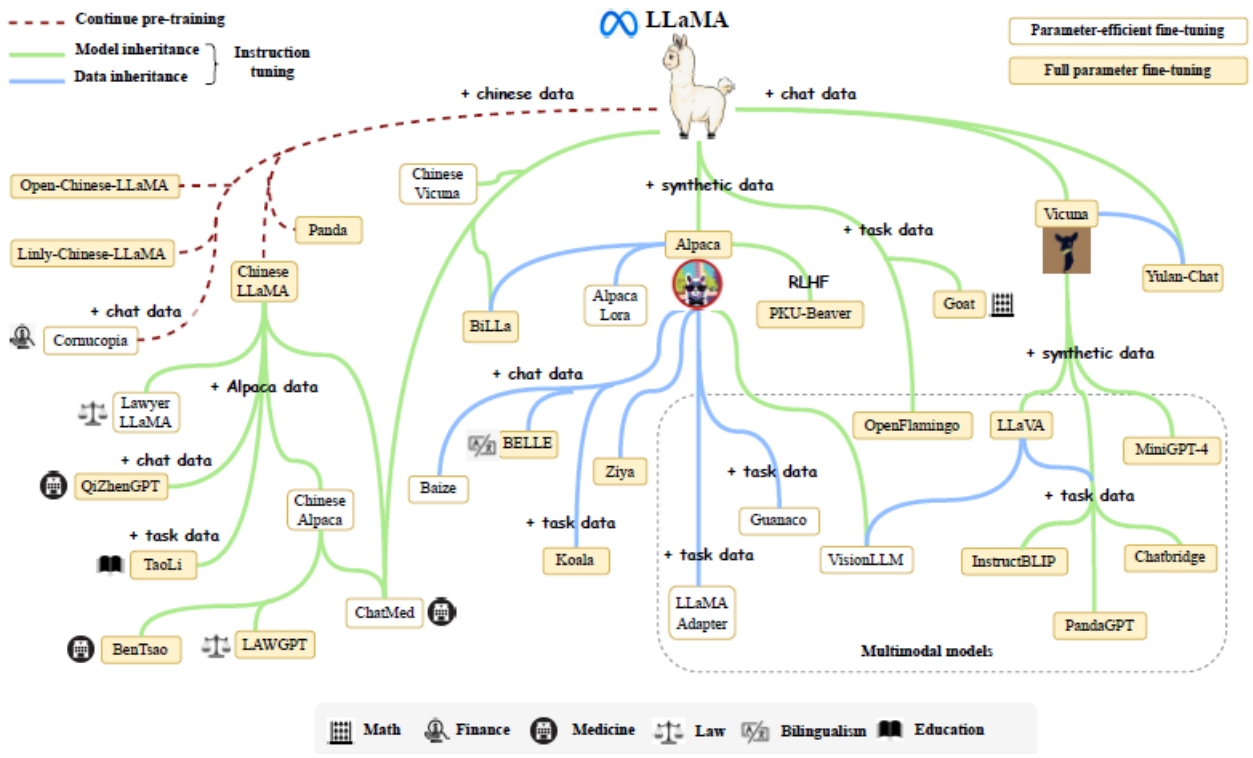


Figure 18 Evolutionary graph of the research work conducted on LLAMA [DR5], a LLM released by Meta as open-source and free for research, and later also for commercial uses. The authors of [DR5] underlined that due to the huge number of new models created from LLAMA, they could not include all the variants in the figure.

4.5 Monetization Strategies

As for its components, the monetization strategy of generative AI is contingent upon the perspective of the actor assessing the system. We propose a comprehensive examination of these monetization strategies in the context of distinct actors:

- AI Model Provider: follows in general two different strategies. The first strategy is to provide the model open source with no license costs. In this case the model integrator will support the infrastructure costs required to host the model. The second strategy is to provide the model as a service (MaaS). In this case, access to models is provide through an API, for which integrators are charged for the number of tokens processed by the model in input to analyse the user request and for the number of contents (number of tokens for language, number of images, etc.) generated in output. The cloud platform hosting the model of the model provider, will charge the model provider for the compute instances used in the cluster by following a pay-per-use model or a reserved model if the customer reserves the compute instances on the long term.
- AI Model Integrator: Application or service providers incorporating generative AI models typically choose the subscription model. In this arrangement, end-users are charged a monthly fee for the application, particularly if it is designed specifically for generative AI. Alternatively, users may pay an additional monthly subscription for added features integrated into a broader product, such as Copilot for Microsoft 365 or Copilot for Github.

As models, AI accelerators, and cloud platforms are central components of all applications and services integrating generative AI features, we think that model providers, GPU providers and cloud providers will likely capture most of this monetization.

4.6 Market Consolidation

Authors of [DR1], recall that today, several ongoing antitrust lawsuits in the US and Europe involve major technology companies like Apple, Google/Alphabet, and Facebook/Meta.

In the market of generative AI, we are once again witnessing the same companies taking the lead in technological advancements, leaving little room for others to compete.

Examples of startup like OpenAI and Mistral AI show that, because of the enormous computing resources required and their related costs, emerging startups struggle to remain independent when they want to scale their models. These are opportunities for technology giants like Microsoft to take advantages of their historical resources and conclude exclusive partnerships.

Big cloud providers like Microsoft, Google or Amazon show a high level of vertical integration:

- they provide the large GPU clusters required for training and provisioning the foundation models,
- they provide their own AI platforms and in complement AI platforms of tiers providers,
- they provide their own foundation models and foundation models of tiers providers,
- some like Google or Microsoft provide the deep learning frameworks on which rely all the code of model training and inference,
- they provide all resources and tools required to develop the applications integrating the generative AI, but also directly some of the applications using generative AI (like the smart assistants or the web browsers).

Highlighted by the authors of [DR1], Google stands out as possibly the company with the most extensive vertical integration within the generative AI market. However, other technology giants such as Microsoft and Amazon are not lagging far behind and are consistently enhancing their level of vertical integration, evident in initiatives like the development of their AI accelerator chips.

We think that technology giants like Google, Microsoft and Amazon with a high level of vertical integration have a clear advantage on the competition. They capture monetization of the generative AI components over almost all layers shown in the Figure 12. They have the financial and technical resources to develop their own models. They can leverage these financial and technical resources to develop partnerships with the actors emerging on the market.

Most of the technology giants already have access to leading-edge technologies, attract the top talents in AI and have built a strong intellectual property portfolio. So, we think that big companies by acquiring or financing smaller companies in generative AI, try essentially to strengthen their existing offering by expanding their product and service offerings and improving their existing AI capabilities.

On January 25th, 2024, the Federal Trade Commission of United States declared to inquiry into Generative AI investments and partnerships[10]. The compulsory orders were sent to Alphabet, Inc., Amazon.com, Inc., Anthropic PBC, Microsoft Corp., and OpenAI, Inc. The FTC declares “Our study will shed light on whether investments and partnerships pursued by dominant companies risk distorting innovation and undermining fair competition." This could also trigger the need for EU to adapt antitrust investigation tools and practices in the near future.

[10] https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships

A. References cited in this report

(1) Exploring Antitrust and Platform Power in Generative AI, 10/07/2023, arXiv:2306.11342v3

(2) https://www.gartner.com/en/topics/generative-ai

(3) A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT, 07/03/2023, arXiv:2303.04226v1

(4) History of the Generative AI, Gilles Legoux, https://medium.com/@glegoux/history-of-the-generative-ai-aa1aa7c63f3c

(5) A survey of Large Language Models, 24/11/2023, arXiv:2303.18223v13

(6) A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?, 21/03/2023, arXiv:2303.11717v1

(7) Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, 27/04/2023, arXiv:2304.13712v2

(8) The Power Of Fine-Tuning In Generative AI, 10/10/2023, Forbes, https://www.forbes.com/sites/forbestechcouncil/2023/10/10/the-power-of-fine-tuning-in-generative-ai/

(9) LLM in a flash: Efficient Large Language Model Inference with Limited Memory, 04/01/2024, Apple Research, arXiv:2312.11514v2

(10) The economic potential of generative AI: The next productivity frontier, June 2023, Mc Kinsey.

(11) Survey of Hallucination in Natural Language Generation, 19/02/2024, Center for Artificial Intelligence Research, Hong Kong University, arXiv:2202.03629v6

(12) From out-of-the-box to tailor-made: Developing and deploying enterprise generative AI tools, Survey by Altman Solon, May 2023, on a panel of 292 senior business leaders and 21 industry experts.

(13) The state of AI in 2020, 17/11/2020, Survey by Mc Kinsey, https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020

(14) Scaling laws for Neural Language Models, 23/01/2020, OpenAI, arXiv:2001.08361v1

(15) Training Compute-Optimal Large Language Models, 29/03/2022, DeepMind, arXiv:2203.15556v1

(16) What Large Models Cost You, 08/09/2023, Forbes, https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/

(17) Generative AI: The Global debate and controversies on use of copyrighted content as training data, 09/03/2023, Anthony Wong, https://unctad.org/news/cstd-dialogue-anthony-wong

(18) A Survey on Bias and Fairness in Machine Learning, 25/01/2022, USC-ISI, arXiv:1908.09635v3

(19) The state of AI in 2023: Generative AI's breakout year, August 2023, Mc Kinsey

(20) Nvidia Chip Shortages Leave AI Startups Scrambling for Computing Power, 24/08/2023, Wired, Nvidia Chip Shortages Leave AI Startups Scrambling for Computing Power | WIRED

(21) The big bottleneck for AI: a shortage of powerful chips, 06/08/2023, CNN Business, The big bottleneck for AI: a shortage of powerful chips | CNN Business

(22) Microsoft warns of service disruptions if it can't get enough A.I. chips for its data centers, 28/07/2023, CNBC, https://www.cnbc.com/2023/07/28/microsoft-annual-report-highlights-importance-of-gpus.html

(23) What's in a Name? Auditing Large Language Models for Race and Gender Bias, 21/02/2024, Standford law school, arXiv:2402.14875v1

(24) Auditing large language models: a three-layered approach, 30/05/2023, Published by Springer, Authors of Oxford, Princeton, Goethe University and Bologna university

(25) Open Source AI: Opportunities and Challenges, 09/02/2024, I. Wladawsky, Linux foundation

(26) The state of generative AI in 7 charts, 02/08/2023, A survey by CBInsights, The generative AI landscape: Top startups, venture capital firms, and more (cbinsights.com)

B. The increased complexity and scale of Generative AI models

The increased complexity of Generative AI models is due at first to the increased complexity of their backbone architecture and to their need to scale up to reach higher performance levels.

Then the number of modalities handled by the models increase again the complexity to overcome the multiple challenges due to the various types of data processed by the model.

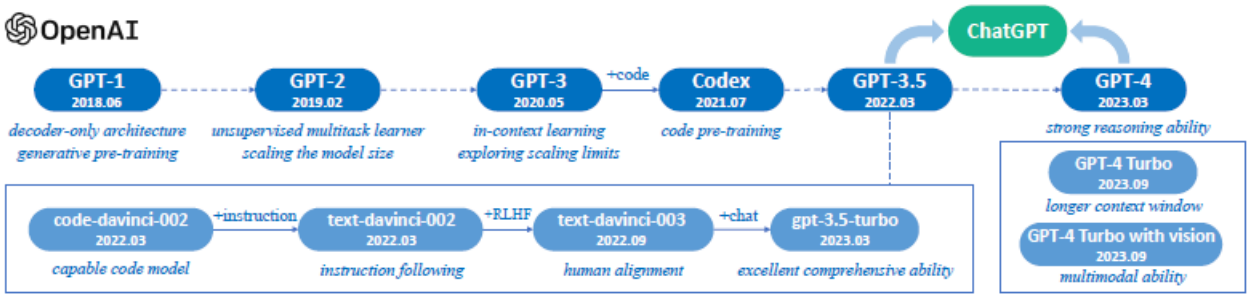


Figure 19 Brief illustration for the technical evolution of GPT-series models [DR5].

A good illustration of this increased complexity is the evolution of GPT-series models by OpenAI (see Figure 19). OpenAI initiated the series with a transformer-based backbone architecture, that has been scale up from 1.5 Billion of parameters in GPT-2 to 175 billion of parameters in GPT-3 to reach better performances. Then the model has been optimized to follow human instructions to integrate ChatGPT in its version 3.5-turbo. Finally, it became multi-modal to generate and ingest text and images, showing also superior ability in reasoning. Its number of parameters is kept secret by OpenAI, but it is estimated to get over 400 billion of parameters.

Researchers now design scaling laws to pave the way to new evolutions of models in term of performances, computation times and costs.

For example, **researchers of OpenAI** [DR14] shown that there exists a power-law relationship between the metric measuring the model performance and the number of model parameters, the amount of compute used during training, and the amount of training data. They **concluded that the model size was the dominant factor to increase the model performances** (see Figure 20 OpenAI law on scaling up model performances.).

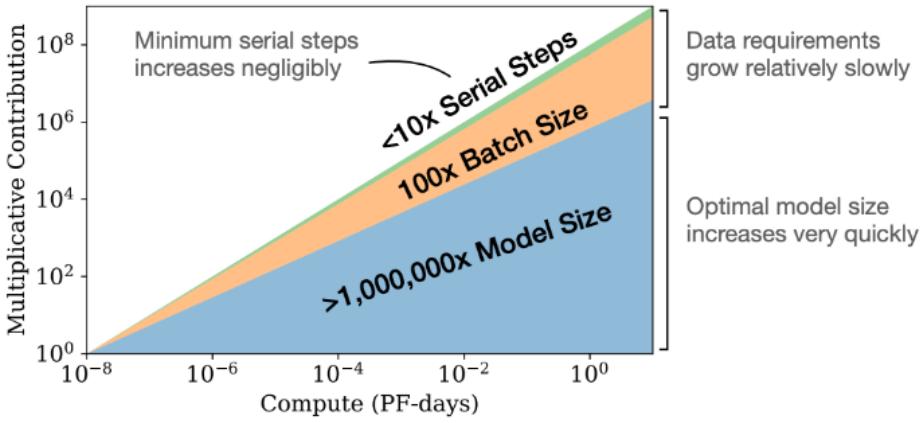
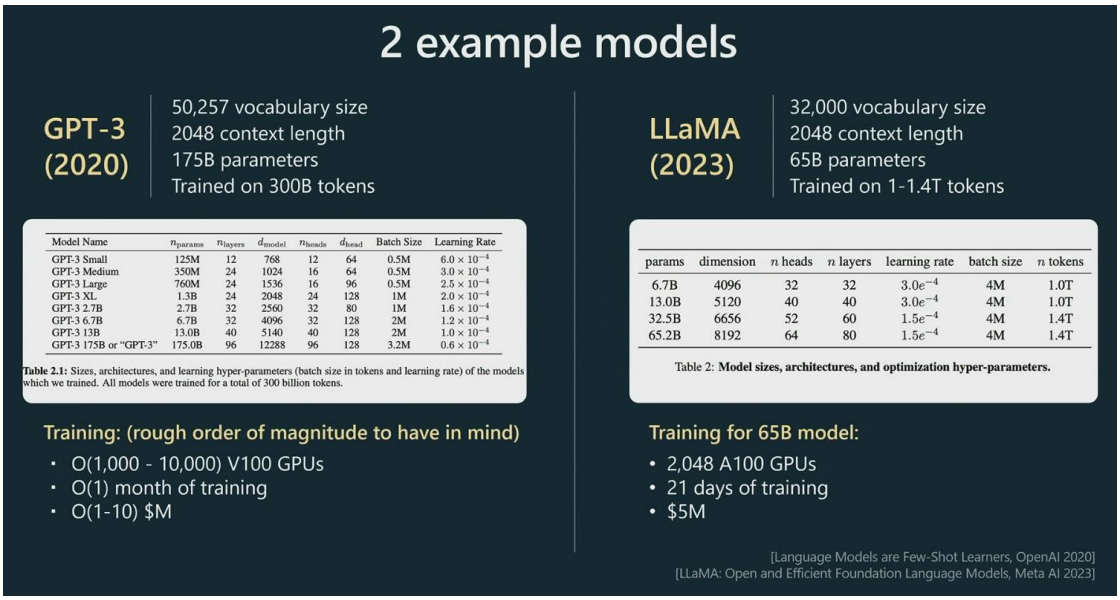


Figure 20 OpenAI law on scaling up model performances.

Large language models like Megatron (by Nvidia), GPT-3 (by OpenAI) and PaLM (by Google) were designed with this approach to effectively utilize increased compute resources. This explains why very large clusters of thousands of high-end GPUs are leveraged during weeks for training these models with costs estimated of several millions of dollars.



Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

Figure 21 Infrastructure requirements and costs for training two iconic of the Large Language Models. ***Source: The state of GPT, May 2023, Microsoft conference by Andrej Karpathy (OpenAI).***

In 2022 **researchers of DeepMind** [DR15], introduced new scaling laws that prioritize the importance of training data for minimizing loss. Unlike previous approaches that emphasized increasing parameters, this paper suggests scaling the number of parameters and training tokens equally. DeepMind initially trained a large model named Gopher with 280 billion parameters. However, they **found that a smaller model named Chinchilla, with 70 billion parameters but trained on four times as much data, performed better within the same compute budget**. Notably, Chinchilla even outperformed larger models like GPT-3.

The model LLAMA2 of Meta was designed by following this last law, this explains why LLAMA2 reaches equivalent performances to GPT-3 but with a reduced number of parameters.

When we compare computation costs for training of GPT-3 and LLAMA2, we see that despite this difference in the approach, training costs of models remain in the same order of magnitude (< 10 million of dollars for GTP-3 compared to 5 million of dollars for LLAMA2). This can be explained by the fact that even if the training of LLAMA2 ran on less epochs of training, one epoch of training includes more iterations as the dataset to explore is much larger.

So what would be the benefits with this second approach? A smaller model requires a smaller infrastructure to serve the model for the same number of users, and so lower the provision costs of LLMs.

Nowadays complexity and scale of generative AI models requires highly knowledgeable and skilled talents, complex and huge infrastructure, where only computing costs for training are estimated to be over several millions or tens of millions of dollars. This is a huge barrier to entry and expansion that many companies cannot overcome to develop foundation models.