Title: Foundation model

URL: https://en.wikipedia.org/wiki/Foundation_model

PageID: 70984276

Categories: Category:Computational fields of study, Category:Computational linguistics, Category:Deep learning, Category:Language modeling, Category:Natural language processing, Category:Unsupervised learning

-----

In artificial intelligence (AI), a foundation model ( FM ), also known as large X model ( LxM ), is a machine learning or deep learning model trained on vast datasets so that it can be applied across a wide range of use cases. [ 1 ] Generative AI applications like large language models (LLM) are common examples of foundation models. [ 1 ]

Building foundation models is often highly resource-intensive, with the most advanced models costing hundreds of millions of dollars to cover the expenses of acquiring, curating, and processing massive datasets, as well as the compute power required for training. [ 2 ] These costs stem from the need for sophisticated infrastructure, extended training times, and advanced hardware, such as GPUs . In contrast, adapting an existing foundation model for a specific task or using it directly is far less costly, as it leverages pre-trained capabilities and typically requires only fine-tuning on smaller, task-specific datasets.

Early examples of foundation models are language models (LMs) like OpenAI's GPT series and Google 's BERT . [ 3 ] [ 4 ] Beyond text, foundation models have been developed across a range of modalities—including DALL-E , Stable diffusion , and Flamingo [ 5 ] for images, MusicGen [ 6 ] and LLark [ 7 ] for music, and RT-2 [ 8 ] for robotic control. Foundation models are also being developed for fields like astronomy, [ 9 ] radiology, [ 10 ] genomics, [ 11 ] coding, [ 12 ] times-series forecasting, [ 13 ] mathematics, [ 14 ] and chemistry. [ 15 ]

## Definitions

The Stanford Institute for Human-Centered Artificial Intelligence's (HAI) Center for Research on Foundation Models (CRFM) coined the term "foundation model" in August 2021 [ 16 ] to mean "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks". [ 17 ] This was based on their observation that preexisting terms, while overlapping, were not adequate, stating that "' (large) language model ' was too narrow given [the] focus is not only language; 'self-supervised model' was too specific to the training objective; and 'pretrained model' suggested that the noteworthy action all happened after 'pretraining." [ 18 ] The term "foundation model" was chosen over "foundational model" [ 19 ] because "foundational" implies that these models provide fundamental principles in a way that "foundation" does not. [ 20 ]

As governments regulate foundation models, new legal definitions have emerged.

In the United States, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence defines a foundation model as "an AI model that is trained on broad data; generally uses self-supervision ; contains at least tens of billions of parameters; is applicable across a wide range of contexts". [ 21 ]

In the United States, the proposed AI Foundation Model Transparency Act of 2023 [ 22 ] by House Representatives Don Beyer (D, VA) and Anna Eshoo (D, CA) defines a foundation model as "an artificial intelligence model trained on broad data, generally uses self supervision, generally contains at least 1,000,000,000 parameters, is applicable across a wide range of contexts, and exhibits, or could be easily modified to exhibit, high levels of performance at tasks that could pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters."

In the European Union, the European Parliament 's negotiated position on the E.U. AI Act defines a foundation model as an "AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks".

In the United Kingdom, the Competition and Markets Authority 's AI Foundation Models: Initial Report [ 1 ] defines foundations model as "a type of AI technology that are trained on vast amounts of data that can be adapted to a wide range of tasks and operations."

The United States's definitions are the only ones to make reference to the size of a foundation model, and differ on magnitude. Beyer and Eshoo's definition also specifies that foundation models must achieve a level of performance as to be a potential danger. In contrast, the E.U. definition requires the model to be designed for generality of output. All definitions agree that foundation models must be trained on a broad range of data with potential applications in many domains.

## History

Technologically, foundation models are built using established machine learning techniques like deep neural networks , transfer learning , and self-supervised learning . Foundation models differ from previous techniques as they are general purpose models that function as a reusable infrastructure, instead of bespoke and one-off task-specific models.

Advances in computer parallelism (e.g., CUDA GPUs ) and new developments in neural network architecture (e.g., Transformers ), and the increased use of training data with minimal supervision all contributed to the rise of foundation models. Foundation models began to materialize as the latest wave of deep learning models in the late 2010s. [ 23 ] Relative to most prior work on deep learning, these language models demonstrated the potential of training on much larger web-sourced datasets using self-supervised objectives (e.g. predicting the next word in a large corpus of text). These approaches, which draw upon earlier works like word2vec and GloVe , deviated from prior supervised approaches that required annotated data (e.g. crowd-sourced labels).

The 2022 releases of Stable Diffusion and ChatGPT (initially powered by the GPT-3.5 model) led to foundation models and generative AI entering widespread public discourse. Further, releases of LLaMA , Llama 2, and Mistral in 2023 contributed to a greater emphasis placed on how foundation models are released with open foundation models garnering a lot of support [ 24 ] and scrutiny. [ 25 ]

## Related concepts

### Frontier models

Certain highly advanced foundation models are termed "frontier models", which have the potential to "possess dangerous capabilities sufficient to pose severe risks to public safety." [ 26 ] These "dangerous capabilities" stem from the accidental or intentional misuse of such models, which in conjunction with their powerful nature can lead to severe harms. As foundation models continue to improve, some AI researchers speculate that almost all next-generation foundation models will be considered frontier models.

Since the concept of dangerous capabilities is inherently subjective, there is no strict designation for what foundation models qualify as frontier models. However, some generally held ideas for sufficiently dangerous capabilities include:

Designing and synthesizing new biological or chemical weapons [ 27 ]

Producing and propagating convincing, tailored disinformation with minimal user instruction [ 28 ]

Harnessing unprecedented offensive cyber capabilities [ 29 ]

Evading human control through deceptive means [ 30 ]

Due to frontier models' unique capabilities, it is difficult to effectively regulate their development and deployment. Because of their emergent nature, new dangerous capabilities can appear on their own in frontier models, both in the development stage and after being deployed. [ 26 ] Additionally, since frontier models continue to adapt after deployment, it remains difficult to mitigate all harms that arise from already-deployed models. If a frontier model happens to be open-source or is released online, the model can also disseminate rapidly, further hampering regulators by creating a lack of accountability.

## General-purpose AI

Due to their adaptability to a wide range of use-cases, foundation models are sometimes considered to be examples of general-purpose AI. In designing the EU AI Act, the European Parliament has stated that a new wave of general-purpose AI technologies shapes the overall AI ecosystem. [ 31 ] The fuller structure of the ecosystem, in addition to the properties of specific general-purpose AI systems, influences the design of AI policy and research. [ 32 ] General-purpose AI systems also often appear in people's everyday lives through applications and tools like ChatGPT or DALL-E .

Government agencies like EU Parliament have identified regulation of general-purpose AI, such as foundation models, to be a high priority. General-purpose AI systems are often characterized by large size, opacity, and potential for emergence, all of which can create unintended harms. Such systems also heavily influence downstream applications, which further exacerbates the need for regulation. In regards to prominent legislation, a number of stakeholders have pushed for the EU AI Act to include restrictions on general-purpose AI systems, all of which would also apply to foundation models.

## World models

In 2018, researchers David Ha and Jürgen Schmidhuber defined world models in the context of reinforcement learning : an agent with a variational autoencoder model V for representing visual observations, a recurrent neural network model M for representing memory, and a linear model C for making decisions. They suggested that agents trained on world models in environments that simulate reality could be applied to real world settings. [ 33 ]

In 2022, Yann LeCun saw a world model (defined by him as a neural network that acts as a mental model for aspects of the world that are seen as relevant) as part of a larger system of cognitive architecture – other neural networks that are analogous to different regions of the brain . In his view, this framework could lead to commonsense reasoning . [ 34 ] [ 35 ]

World models are trained on a variety of data modalities, including text, images, audio and video, and have been applied to video generation . [ 36 ]

TechCrunch noted that world models could use more data than large language models and would require significantly more computational power (including the use of thousands of GPUs for training and inference). [ 35 ] [ 36 ] It also noted the risk of hallucinations , coverage bias and algorithmic bias . [ 36 ]

TechCrunch saw Sora as an example of a world model, [ 36 ] while in January 2025, Nvidia released its own set of world models. [ 37 ] [ 38 ] The South China Morning Post wrote that Manycore Tech was another example of companies aiming to build a world model, viewing their work as an example of spatial intelligence . [ 39 ] In May 2025, Mohamed bin Zayed University of Artificial Intelligence released a world model for building simulations to test AI agents . [ 40 ]

Google DeepMind has also released two world models in two-dimensional space and three-dimensional space , respectively, that were trained on video data, with Google claiming that the latter can be a training environment for AI agents. [ 41 ] [ 42 ]

Fei-Fei Li views world models as applying to robotics and creative works . Due to the complexity of these models, she advocates for more complex strategies in data acquisition , data engineering , data processing , and synthesizing data . [ 43 ] She co-founded a startup on building world models, which, as of 2024, planned to do so in three phases: incorporating an understanding of three-dimensional space along with time; support for augmented reality ; and support for robotics. [ 44 ]

World models are intended for use in interactive media and environment simulation. Creative professionals have expressed concern that world models could disrupt jobs in their industries. [ 45 ] Wired compared world models to the metaverse , [ 44 ] while Business Insider noted possible military applications . [ 43 ]

## Technical details

### Modeling

For a foundation model to effectively generalize, it must acquire rich representations of the training data. As a result, expressive model architectures that efficiently process large-scale data are often preferred in building foundation models. [ 17 ] Currently, the Transformer architecture is the de facto choice for building foundation models across a range of modalities. [ 46 ]

### Training

Foundation models are built by optimizing a training objective(s), which is a mathematical function that determines how model parameters are updated based on model predictions on training data. [ 47 ] Language models are often trained with a next-tokens prediction objective, which refers to the extent at which the model is able to predict the next token in a sequence. Image models are commonly trained with contrastive learning or diffusion training objectives. For contrastive learning, images are randomly augmented before being evaluated on the resulting similarity of the model's representations. For diffusion models, images are noised and the model learns to gradually de-noise via the objective. Multimodal training objectives also exist, with some separating images and text during training, while others examine them concurrently. [ 48 ] In general, the training objectives for foundation models promote the learning of broadly useful representations of data.

With the rise of foundation models and the larger datasets that power them, a training objective must be able to parse through internet-scale data for meaningful data points. Additionally, since foundation models are designed to solve a general range of tasks, training objectives ought to be domain complete , or able to solve a broad set of downstream capabilities within the given domain. Lastly, foundation model training objectives should seek to scale well and be computationally efficient. With model size and compute power both being relevant constraints, a training objective must be able to overcome such bottlenecks.

### Data

Foundation models are trained on a large quantity of data, working under the maxim "the more data, the better." [ 49 ] Performance evaluation does show that more data generally leads to better performance, but other issues arise as data quantity grows. Tasks like managing the dataset, integrating data across new applications, ensuring adherence to data licenses, and maintaining data quality all become more difficult as data size grows. The specific demands of foundation models have only exacerbated such issues, as it remains the norm for large foundation models to use public web-scraped data. Foundation models include also search engines data and SEO meta tags data. Public web data remains a plentiful resource, but it also demands stringent moderation and data processing from foundation model developers before it can be successfully integrated into the training pipeline. [ 50 ]

Training foundation models often runs the risk of violating user privacy, as private data can be disclosed, collected, or used in ways beyond the stated scope. Even if no private data is leaked, models can still inadvertently compromise security through learned behavior in the resulting foundation model. [ 51 ] Data quality is another key point, as web-scraped data frequently contains biased, duplicate, and toxic material. Once foundation models are deployed, ensuring high-quality data is still an issue, as undesirable behavior can still emerge from small subsets of data.

### Systems

The size of foundation models also brings about issues with the computer systems they run on. The average foundation model is too large to be run within a single accelerator's memory and the initial training process requires an expensive amount of resources. [ 52 ] Such issues are predicted to further exacerbate in future as foundation models grow to new heights. Due to this constraint, researchers have begun looking into compressing model size through tight model inference.

GPUs are the most common choice of compute hardware for machine learning, due to high memory storage and strong power. Typical foundation model training requires many GPUs, all connected in parallel with fast interconnects. Acquiring a sufficient amount of GPUs of requisite compute efficiency is a challenge for many foundation model developers, one that has led to an increasing dilemma in the field. Larger models require greater compute power, but often at the cost of improved compute efficiency. Since training remains time-consuming and expensive, the tradeoff between compute power and compute efficiency has led only a few select companies to afford the production costs for large, state of the art foundation models. Some techniques like compression and distillation can make inference more affordable, but they fail to completely shore up this weakness.

## Scaling

The accuracy and capabilities of foundation models often scale predictably with the size of the model and the amount of the training data. Specifically, scaling laws have been discovered, which are data-based empirical trends that relate resources (data, model size, compute usage) to model capabilities. Particularly, a model's scale is defined by compute, dataset size, and the number of parameters, all of which exhibit a power-law relationship with end performance.

However, broken scaling laws [ 53 ] have been discovered in which this relationship smoothly transitions (at points referred to as break(s) ) from a power law with one exponent to a power law with another (different) exponent. When one does not collect any points near (or after) the break(s), it can be difficult to obtain an accurate extrapolation.

## Adaptation

Foundation models are inherently multi-purpose: to use these model for a specific use case requires some form of adaptation. At a minimum, models need to be adapted to perform the task of interest (task specification), but often better performance can be achieved by more extensive adaptation to the domain of interest (domain specialization).

A variety of methods (e.g. prompting , in-context learning , fine-tuning , LoRA ) provide different tradeoffs between the costs of adaptation and the extent to which models are specialized. Some major facets to consider when adapting a foundation model are compute budget and data availability. Foundation models can be very large, up to trillions of parameters in size, so adapting the entirety of a foundation model can be computationally expensive. Therefore, developers sometimes adapt only the last neural layer or only the bias vectors to save time and space. [ 54 ] For particularly niche applications, specific data may also not be available to adapt the foundation model sufficiently. In such circumstances, data must be manually labeled, which is costly and can demand expert knowledge.

## Evaluation

Evaluation is a key part of developing foundation models. Not only does evaluation allow for tracking progress of high-performance models, it also creates benchmarks for future model development. Stakeholders rely on evaluations to understand model behaviors and gain insight into their various attributes. Traditionally, foundation models are evaluated relative to each other through standardized task benchmarks like MMLU , [ 55 ] MMMU, [ 56 ] HumanEval, [ 57 ] and GSM8K. [ 58 ] Given that foundation models are multi-purpose, increasingly meta-benchmarks are developed that aggregate different underlying benchmarks. Examples include LM-Harness, [ 59 ] BIG-Bench, [ 60 ] HELM, [ 61 ] OpenLLM Leaderboard, [ 62 ] DecodingTrust, [ 63 ] and HEIM. [ 64 ]

Since foundation models' utility depends on their own general capabilities and the performance of fine-tuned applications, evaluation must cover both metrics. Proper evaluation examines both a foundation model's downstream applications in aggregate and the direct properties the foundation model holds. To ensure further equity in evaluation, certain existing evaluation frameworks account for all adaptation resources, which leads to more informed analyses for the benefit of all stakeholders. [ 65 ]

## Supply chain

Foundation models' general capabilities allow them to fulfill a unique role in the AI ecosystem, [ 66 ] fueled by many upstream and downstream technologies. [ 1 ] Training a foundation model requires several resources (e.g. data, compute, labor, hardware, code), with foundation models often involving immense amounts of data and compute (also referred to as computational power). Due to foundation models' large development costs and inexpensive adaptation requirements, the AI landscape has shifted to a small subset of AI companies making foundation models for downstream adaptation. [ 67 ] Thus, most foundation model companies outsource this step to specialized data providers (e.g. Scale AI , [ 68 ] Surge [ 69 ] ) and compute providers (e.g. Amazon Web Services , Google Cloud , Microsoft Azure ).

The foundation model developer itself will then take the data and use the supplied compute to actually train the foundation model. After the foundation model is completely built, much of the data and labor requirements abate. In this development process, hardware and compute are the most necessary, and also the most exclusive resources. To train larger and more complex AI, a sufficient amount of compute is key. However, compute is consolidated in the hands of a few, select entities, which most foundation model developers depend on. As such, the foundation model pipeline is concentrated heavily around these providers. Compute is also costly; in 2023, AI companies spent more than 80% of total capital on compute resources. [ 71 ]

Foundation models require a large amount of general data to power their capabilities. Early foundation models scraped from subsets of the internet to provide this data information. As the size and scope of foundation models grows, larger quantities of internet scraping becomes necessary, resulting in higher likelihoods of biased or toxic data. This toxic or biased data can disproportionately harm marginalized groups and exacerbate existing prejudices. [ 72 ]

To address this issue of low-quality data that arose with unsupervised training, some foundation model developers have turned to manual filtering. This practice, known as data labor, comes with its own host of issues. [ 73 ] Such manual data detoxification is often outsourced to reduce labor costs, with some workers making less than $2 per hour. [ 74 ]

The foundation model will then be hosted online either via the developer or via an external organization. Once released, other parties can create applications based on the foundation model, whether through fine-tuning or wholly new purposes. People can then access these applications to serve their various means, allowing one foundation model to power and reach a wide audience.

## Release strategies

After a foundation model is built, it can be released in one of many ways. There are many facets to a release: the asset itself, who has access, how access changes over time, and the conditions on use. [ 75 ] All these factors contribute to how a foundation model will affect downstream applications. [ 76 ] In particular, the two most common forms of foundation model release are through APIs and direct model downloads.

When a model is released via an API , users can query the model and receive responses, but cannot directly access the model itself. Comparatively, the model could be directly downloadable for users to access and modify. Both release strategies are often classified as an open release. The exact definition of an open release is disputed, but widely accepted requirements are provided by the Open Source Initiative .

Some open foundation models are: PaLM 2 , Llama 2 , Granite , and Mistral . While open foundation models can further research and development more easily, they are also more susceptible to misuse. Open foundation models can be downloaded by anyone, and particularly powerful models can be fine-tuned to intentionally or unintentionally cause harm. [ citation needed ]

During a closed release, the foundation model cannot be accessed by the public, but is used internally by an organization. Such releases are considered safer, but offer no additional value to the research community or the public at large.

Some foundation models like Google DeepMind 's Flamingo [ 77 ] are fully closed, meaning they are available only to the model developer; others, such as OpenAI 's GPT-4 , are limited access, available to the public but only as a black box ; and still others, such as Meta 's Llama 2 are open, with broadly available model weights enabling downstream modification and scrutiny.

## References

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora

Firefly

Flux

GPT Image 1

Ideogram

Imagen

Midjourney

Qwen-Image

Recraft

Seedream

Stable Diffusion

Dream Machine

Hailuo AI

Kling

Midjourney Video

Runway Gen

Seedance

Sora

Veo

Wan

15.ai

Eleven

MiniMax Speech 2.5

WaveNet

Eleven Music

Endel

Lyria

Riffusion

Suno AI

Udio

Agentforce

AutoGLM

AutoGPT

ChatGPT Agent

Devin AI

Manus

OpenAI Codex

Operator

Replit Agent

01.AI

Aleph Alpha

Anthropic

Baichuan

Canva

Cognition AI

Cohere

Contextual AI

DeepSeek

ElevenLabs

Google DeepMind

HeyGen

Hugging Face

Inflection AI

Krikey AI

Kuaishou

Luma Labs

Meta AI

MiniMax

Mistral AI

Moonshot AI

OpenAI

Perplexity AI

Runway

Safe Superintelligence

Salesforce

Scale AI

SoundHound

Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category