Deep Learning and Neural Networks

Demetrio Labate

April 20, 2026

Large Language Models (LLMs)

LLM	Developer	Multimodal?	Reasoning?	Access
GPT-5.4	OpenAI	Yes	Yes	API, Chatbot
gpt-oss	OpenAI	No	Yes	Open
Gemini	Google	Yes	No	API, Chatbot
Gemma	Google	No	No	Open
Llama	Meta	Yes	No	Open
R1	DeepSeek	No	Yes	Open, API, Chatbot
V3.2	DeepSeek	No	Yes	Open, API, Chatbot
Claude	Anthropic	Yes	Yes	API, Chatbot
Command	Cohere	No	Yes	API
Nova	Amazon	Yes	No	API
Mistral	Mistral	No	Yes	API, Chatbot, Open weight
Qwen	Alibaba Cloud	No	Yes	Open, API, Chatbot
GLM-5	Z.ai	No	Yes	Open, API, Chatbot
Kimi K2.5	Moonshot AI	No	Yes	Open, API, Chatbot
MiniMax M2.5	MiniMax	Yes	Yes	Open, API, chatbot
MiMo-V2-Flash	Xiaomi	No	Yes	Open, API
Phi	Microsoft	Yes	Yes	Open
Grok	xAI	Yes	Yes	API, Chatbot

Large Language Models (LLMs)

There are three main categories of LLMs: proprietary, open, and open source.

Proprietary models like GPT-5.4 and Claude Opus 4.7 are some of the most popular and powerful models available, but they're developed and operated by private companies.

The source code, training strategies, model weights, and even details like the number of parameters they have are all kept secret.

The only way to access these models is through a chatbot or app built with them, or through an API.
You cannot run GPT-5.4 on your own server.

Large Language Models (LLMs)

Open and open source models are more freely available.

You can download Llama 4, gpt-oss-20b, Gemma 3, and DeepSeek V3 from Hugging Face and other model platforms and run them on your own devices—and even re-train them with your own data to create your own model.
They are also available from multiple third-party API providers so developers can build their own chatbots and apps on top of them.

So what is the difference between open and open source?

Open source licenses are fully permissive. Mostly, you have to agree to make anything you build with it open source as well and give attribution to the original developers.

Open licenses are still permissive, but have some additional limits. For example, Llama 4's license allows commercial use up to 700 million monthly users and blocks certain uses. . Similarly, Gemma 3's prohibited use policy, among other things, bans "facilitating or encouraging users to commit any type of crimes."

Large Language Models (LLMs)

In early 2026, here are the primary LLM models.

Proprietary & Leading Models (Closed Source)

- OpenAI GPT-5.x/4o
- Anthropic Claude 3.5 Sonnet / 4 Opus
- Google Gemini 3.x/Flash
- Cohere Command R+.

Open-Weight Models

- Meta Llama 3/4:
- Mistral (Large/Small/NeMo):
- Qwen 3 (Alibaba):
- Gemma 3 (Google):

Leading LLM Model Developers

OpenAI:
OpenAI's Generative Pre-trained Transformer (GPT) models kickstarted the latest AI hype cycle with the release of ChatGPT in November 2022

GPT-5.4, the company's latest flagship model (as of April 2026), is both multimodal and capable of reasoning.

It aims to combine all the features of GPT-4o, o3, and Codex (OpenAI's general purpose, reasoning, and coding models) into a single general purpose model.

GPT-5.4 is available through ChatGPT and as an API for developers.

As a reasoning model, the options available through the API can get complicated as you have to set how much effort the model can expend to work through difficult problems as well as select whether to use the full model, or the nano or mini versions.

Leading LLM Model Developers

OpenAI also has a state-of-the-art open LLM called **gpt-oss**

gpt-oss-20b and gpt-oss-120b are both open reasoning models that were trained using the same techniques as OpenAI's other models like o3 and GPT-4o.

Parameters: 21 billion, 117 billion (as mixtures-of-experts)

Anyone can download, fine-tune, and use these models for almost any purpose (though OpenAI has taken steps to limit the ways they can be used for malicious purposes or to generate harmful information). This is a big deal because since 2019, all of OpenAI's models have been proprietary.

Introducing gpt-oss

Leading LLM Model Developers

Anthropic:
It is considered a top competitor in enterprise AI, with a focus on safety, long-context understanding, and reliability.

It uses the Claude family of models.
Its three hybrid reasoning models - Claude Sonnet 4.6, Claude 4.5 Haiku, and Claude Opus 4.7 — are designed to be helpful, honest, harmless, and safe for enterprise customers to use.

Claude Sonnet 4.6 is now considered one of the best AI coding models available.

Like all the other proprietary LLMs, Claude is only available as an API or through its official chatbot and other products, though it can be further trained on your data and fine-tuned.

Leading LLM Model Developers

Google DeepMind:
It is known for its Gemini model family (Gemini 3 Pro/Flash).

These models are designed for high-end multimodal tasks, long-context windows (1M+ tokens), and deep integration into the Google Cloud ecosystem.

Google Gemma is a family of open AI models from Google based on the same research and technology it used to develop Gemini.

The latest version, Gemma 3, is available in five sizes: 270 million, 1 billion, 4 billion, 12 billion, and 27 billion parameters.

There is also a version called Gemma 3n designed for mobile architectures.

Leading LLM Model Developers

Meta AI:
It leads in open-weight models with the Llama series.
Parameters: 109 billion, 400 billion, 2 trillion (all as mixtures-of-experts)

The newest Llama 4 models (including Scout, Maverick, and Behemoth [in preview]) are multimodal and use a mixture-of-experts structure.
Scout has a 10M context window, which is bigger than anything else available at the moment.

Developers widely use these models for customization and private, cost-efficient deployment.

You can download the source code yourself from GitHub

Leading LLM Model Developers

DeepSeek:
DeepSeek R1 caused a major stir when it was launched, as it was the first state-of-the-art reasoning model developed by a Chinese tech company.

It was created on more limited computer hardware with a far smaller budget and released as an open model.

DeepSeek-R1 is an open model accessible via chatbot and API
Parameters: 671 billion (as a mixture-of-experts)

R1 has been continually updated since launch, though it has been somewhat deprecated in favor of DeepSeek V3.2.

Leading LLM Model Developers

Mistral AI:
This prominent European AI company focuses on open-source and efficient models suitable for various business applications.

Especially known for efficiency and strong performance in smaller sizes (3B, 8B, 14B).

Mistral 3 is its latest non-reasoning frontier model; it uses a mixture-of-experts architecture with 41 billion active and 675 billion total parameters.

Mistral also has the Magistral models that support reasoning, and the Ministral models with 3 billion, 8 billion, and 14 billion parameters.

Leading LLM Model Developers

xAI:
Elon Musk's company is developing the Grok series.

Grok, an AI model and chatbot trained on data from X (formerly Twitter), originally was not very competitive.

Grok 4, however, offers state-of-the-art performance and reasoning abilities. There's also a smaller version available called Grok 4 Fast.

The focus is on real-time data access and future applications in math, physics, and robotics.

Leading LLM Model Developers

Cohere:
It specializes in enterprise-ready models (Command) and is optimized for retrieval augmented generation (RAG) so that organizations can have the model respond accurately to specific queries from employees and customers.

There is often a focus on data privacy.

Companies like Oracle, Accenture, Notion, and Salesforce use Cohere's models

Beyond LLM Models

- The industry is hitting a plateau with language-only reasoning.
- One of the most notable trends is the **shift from Large Language Models (LLMs) to World Models**.
- This involves moving from predicting the next token to simulating physical reality, causality, and spatial environments.
- While LLMs handle language and text, World Models understand "how the world works," enabling AI to plan, reason, and act in 3D space

World Models

In recent years a number of prominent AI researchers have turned toward world models.

• 2023: Yann LeCun championed world models to overcome Large Language Model (LLM) limitations, arguing they lack true physical understanding.

This led to **joint-embedding predictive architectures (JEPA)**, focusing on representing the world rather than pixel-perfect reconstruction.

Joint-Embedding Predictive Architecture (JEPA) is a self-supervised learning approach that learns abstract, high-level representations of data by predicting missing or future information in latent space rather than reconstructing raw pixels or text tokens. It uses two encoders to create embeddings (representations) of input data, aiming to align these representations semantically rather than just matching low-level patterns.

World Models

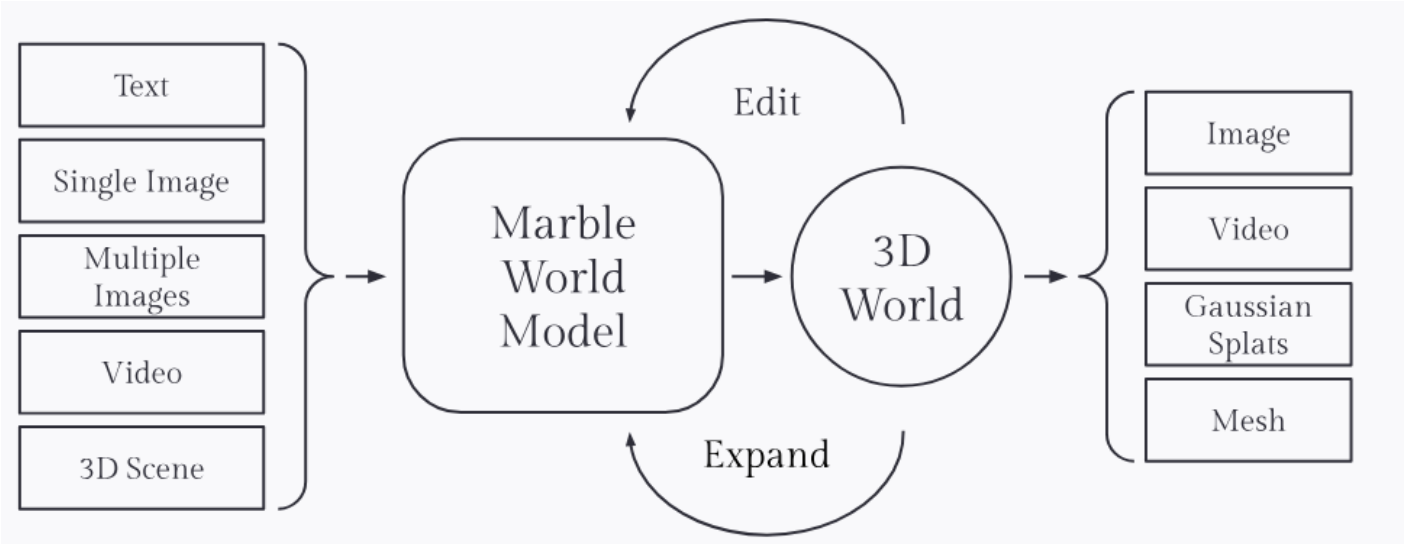
• 2024: **OpenAI** released a large-scale world simulators, called **Sora**, which generates video to simulate physical interactions, alongside robotics initiatives focused on spatial intelligence and 3D environment modeling. **Sora 2** was released in 2025 to focus on superior motion physics, 3D consistency, and multi-modal generation.

The core aim was to create "world models"-AI systems that understand the rules of the physical world (gravity, object permanence) to simulate reality.

However, by March 2026, OpenAI discontinued the Sora app and API to redirect resources towards new language models and robotics.
High Costs and Low Revenue. Competition (Google Genie)

World Models

• 2024 Fei Fei Li founded **World Labs**, which launched its **Marble software** in 2025 to create 3D worlds from "text, images, video, or coarse 3D layouts"



Marble is a world model that can now create 3D worlds from a wide variety of input types, and lets users iteratively edit or expand worlds

Marble

World Models

• 2025 **Google Genie 3** is introduced as a "foundation world model" that generates interactive 2D environments from text, sketches, or images.

Unlike video generators, it creates explorable, playable worlds in real-time, allowing users to move and interact with objects.

It learns physics and logic to enable, for example, creating a playable game from a simple drawing

Interactive Worlds: It does not just create a video; it allows for real-time interaction, letting users play within the generated scene.

World Memory: Maintains consistency within the environment, ensuring objects and scenes remain consistent as the user moves.

Genie3

World Models

- 2026: **Advanced Machine Intelligence Labs (AMI Labs)**, founded by Yann LeCun following his departure from Meta, has closed a $1.03 billion seed round at a $3.5 billion pre-money valuation – Europe's largest seed round on record.

The company is building world models based on LeCun's Joint Embedding Predictive Architecture (JEPA).

Vision: Generative models, which attempt to predict the future state of the world in high-dimensional detail (whether pixel-by-pixel or token-by-token), are necessarily imprecise, because much of what happens in the real world is inherently unpredictable at that level of granularity. JEPA instead trains models to make predictions in abstract representation space – learning what matters about how the world changes, rather than attempting to reconstruct its surface appearance. This approach is better suited to the demands of robotics, industrial process control, wearables, and healthcare, where limitations of LLMs are most consequential.

World Models

From Video Generation to World Model CVPR 2025 Tutorial

Physics-Grounded World Models: Generation, Interaction, and Evaluation