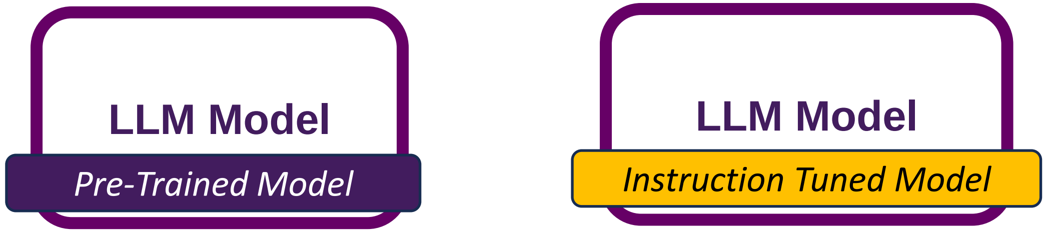
LLM Models

What are LLMs? How do they work?

Intuitive understanding for busy people



LLM Model

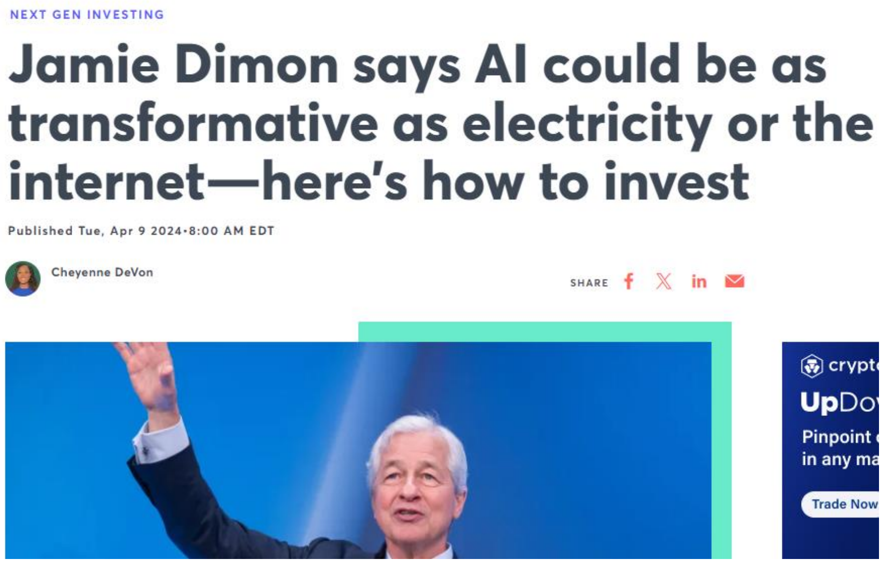
Pre-Trained Model

LLM Model

Instruction Tuned Model

Generative Artificial Intelligence

Generative AI is truly ***revolutionary*** technology. It is transforming the way we interact with technology. We are in a middle of a paradigm shift where for the first-time computers can understand humans via natural language and respond intelligently.



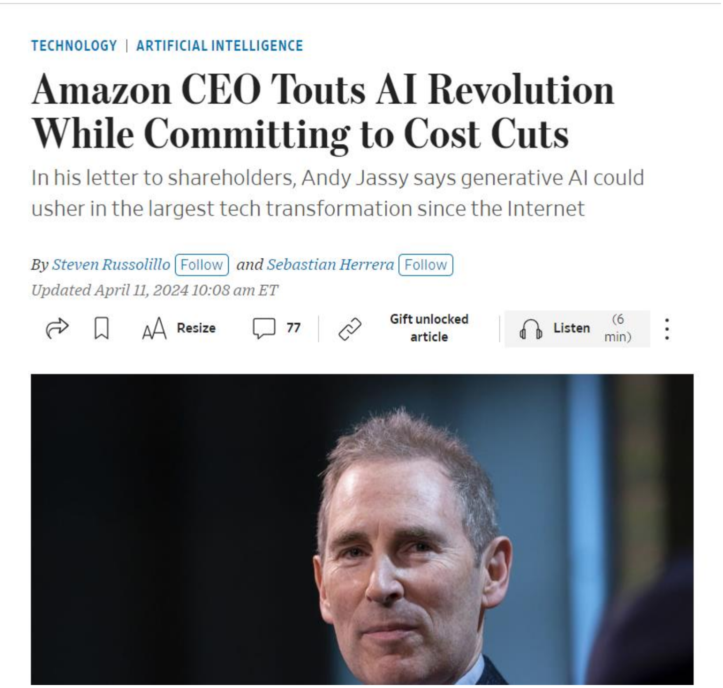
NEXT GEN INVESTING

Jamie Dimon says AI could be as transformative as electricity or the internet—here's how to invest

Published Tue, Apr 9 2024·8:00 AM EDT

Cheyenne DeVon

Source: CNBC



TECHNOLOGY | ARTIFICIAL INTELLIGENCE

Amazon CEO Touts AI Revolution While Committing to Cost Cuts

In his letter to shareholders, Andy Jassy says generative AI could usher in the largest tech transformation since the Internet

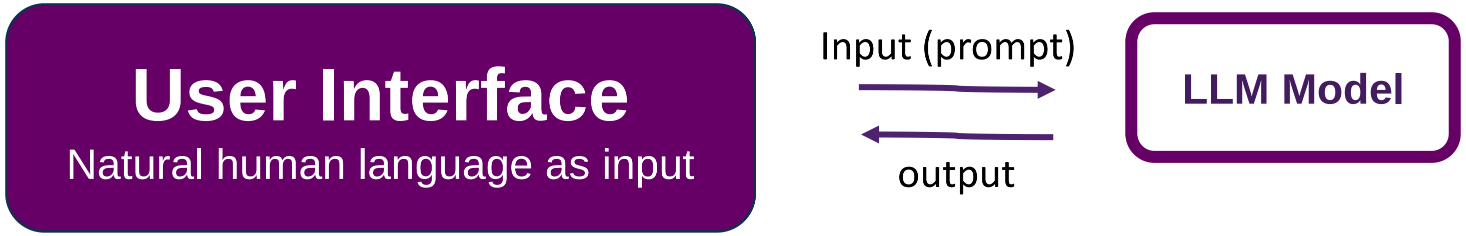
By Steven Russolillo Follow and Sebastian Herrera Follow

Updated April 11, 2024 10:08 am ET

Source: WSJ

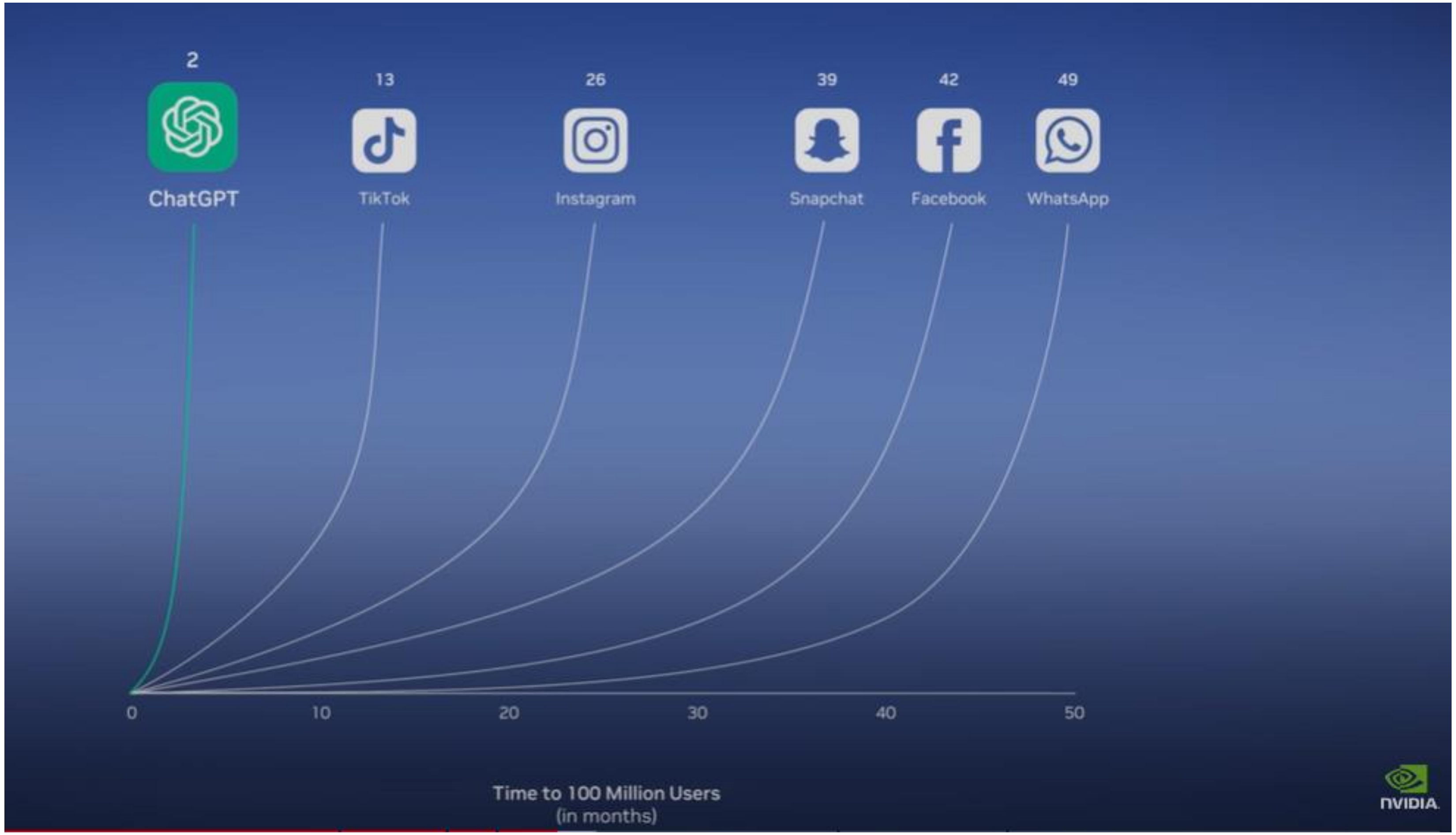
Paradigm Shift : Natural Language

For the first time, we have a universal UI (User Interface). LLMs, can understand understand human natural language and can respond intelligently using natural language.



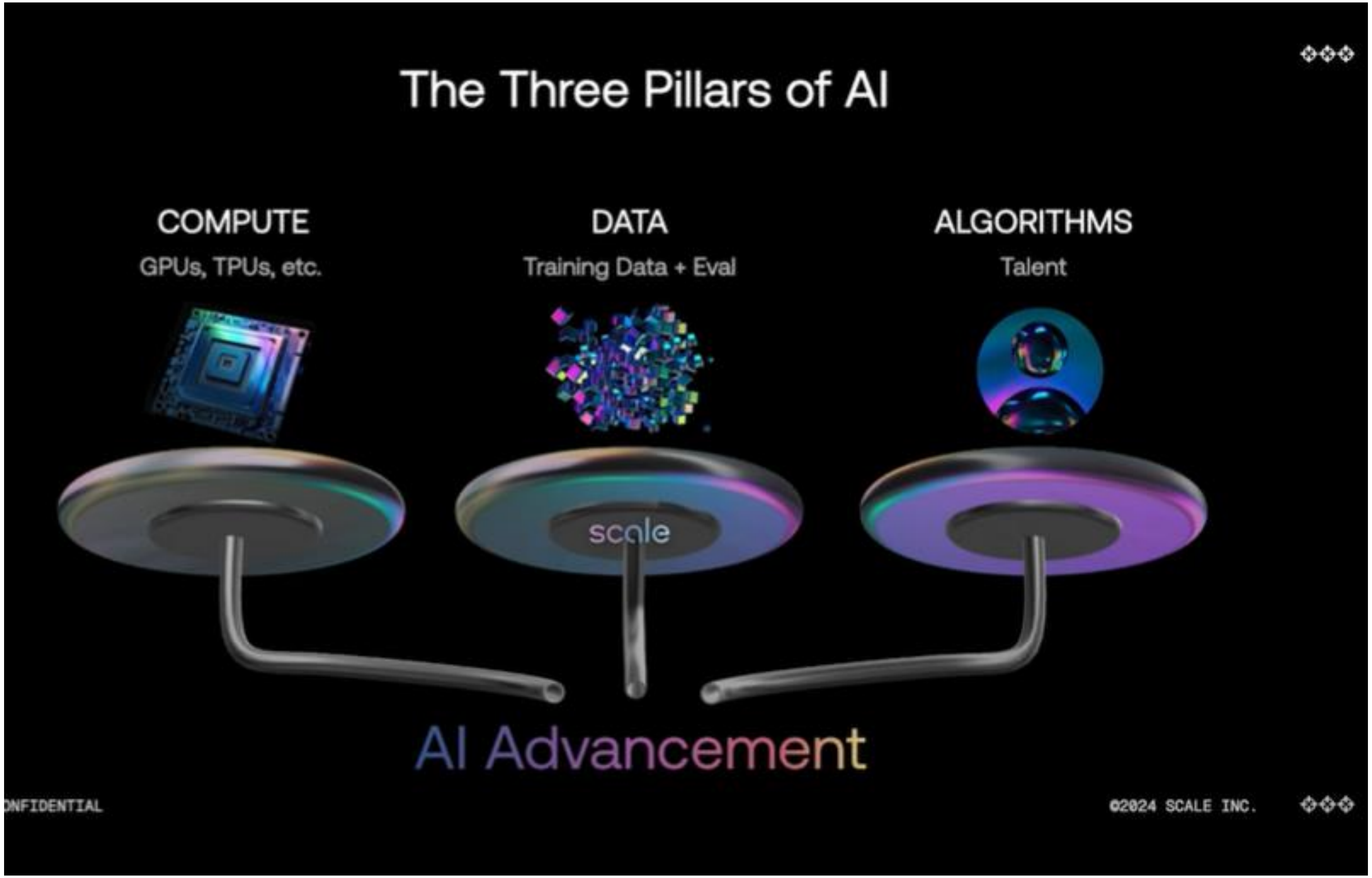
ChatGPT is the fastest growing application in human history.

That is because we use human natural language to interact with it.



Source: Nvidia

Generative Artificial Intelligence



The Three Pillars of AI

COMPUTE
GPUs, TPUs, etc.

DATA
Training Data + Eval

ALGORITHMS
Talent

scale

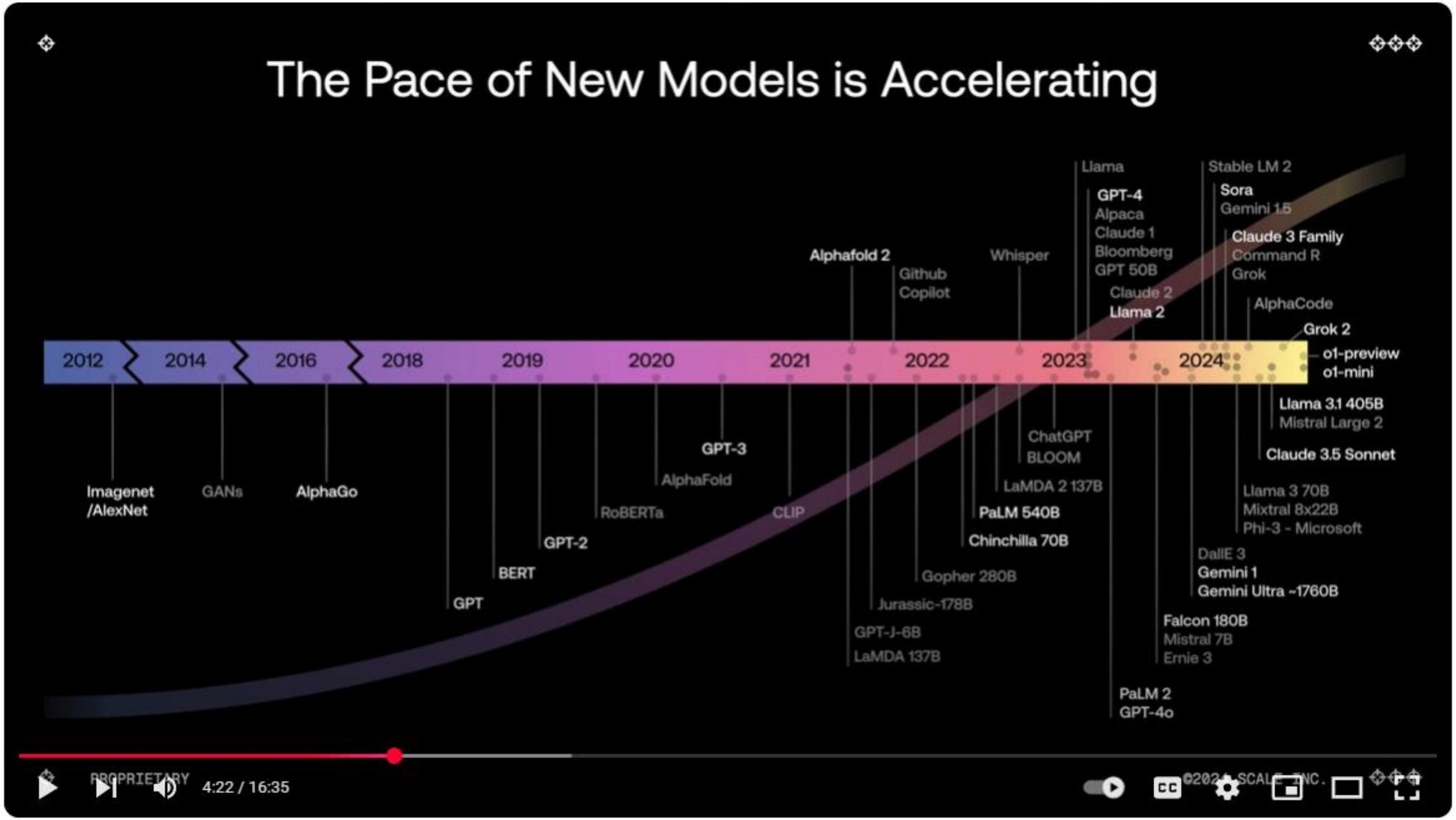
AI Advancement

ONFIDENTIAL

©2024 SCALE INC.

Scale AI Leadership Summit 2024: Alexandr Wang Opening Keynote
https://www.youtube.com/watch?v=eRYP2arKkk0

Generative Artificial Intelligence



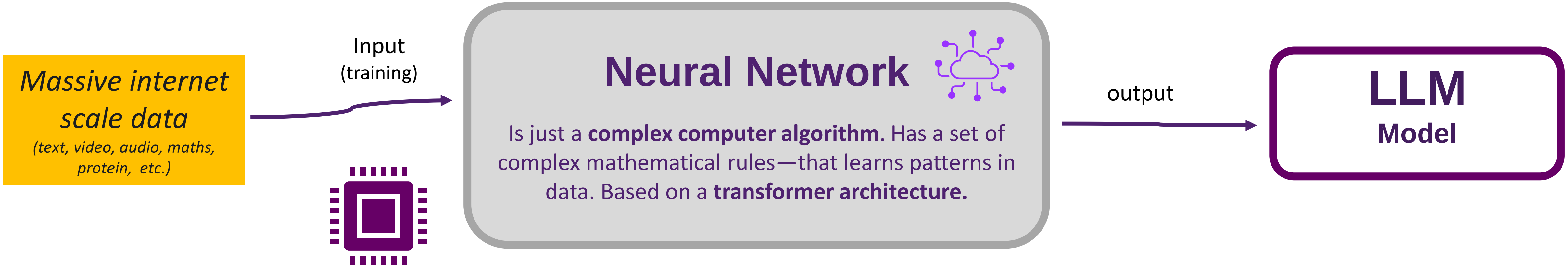
The Pace of New Models is Accelerating

Scale AI Leadership Summit 2024: Alexandr Wang Opening Keynote

https://www.youtube.com/watch?v=eRYP2arKkk0

Large Language Models

Given vast amount of data+compute, an algorithm (called a **neural network**) can programm itself to develop a deep understanding of patterns and meaning in the data on its own. This discipline is called **deep learning**. Once trained, LLMs can use this understanding to generate human like responses when **<u>prompted</u>** using natural language.



Massive internet scale data

(text, video, audio, maths, protein, etc.)

Input (training)

Neural Network

Is just a **complex computer algorithm**. Has a set of complex mathematical rules—that learns patterns in data. Based on a **transformer architecture.**

output

LLM

Model

- Text only models work with text as input & output
- Multimodal models can handle text, image, audio or video as input & output

Transformer Architecture (jargon buster)

- A transformer architecture is a type of algorithm (called a neural network) designed to process sequential data, like sentences, all at once rather than one word at a time.
- It uses a mechanism called **attention**, which helps the model focus on the most important words in a sentence to understand context and meaning.
- This design makes transformers highly efficient and effective at capturing relationships between words, even in long and complex sentences.

Neural Network

Is just a **complex computer algorithm**. Has a set of complex mathematical rules—that learns patterns in data. Based on a **transformer architecture.**

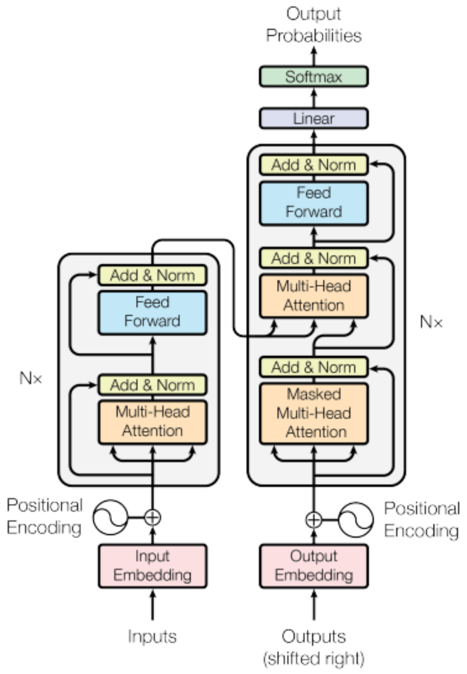


Figure 1: The Transformer - model architecture.

Source: Attention is all you need paper: at https://arxiv.org/pdf/1706.03762

What is inside a LLM model

LLM
Model

- Is a complex computer algorithm. Generally called a “neural network” within the technical community
- This neural network has an architecture called “transformer architecture” : or a set of defined complex mathematical rules—that has capability to learn patterns in data.
- Is a collection of few files. You can even download these files to your PC. The number & type vary based on framework (like TensorFlow or PyTorch). These include
  - Parameter files
  - Configuration/Setup files
  - Runtime files
- Has a "vocabulary size“, which refers to the total number of unique tokens (words, characters, or subwords) that the model recognizes and uses to represent and process text
- It has a number of Layers. Layers can be though of steps in the process of transforming input into output.
- The context window of a LLM refers to the maximum amount of input text (in terms of tokens) that can be sent to the model when generating a response
- The parameters of a LLM can be thought of as variables. The parameter size of a LLM refers to the total number of learnable variables (weights and biases) within the model. A larger parameter size generally means the model can capture more complex patterns and nuances in language, making it more powerful but also requiring more computational resources. For example, GPT-3 has 175 billion parameters, enabling it to generate highly sophisticated and human-like text.
- The process of invoking a LLM is called “inferencing”.

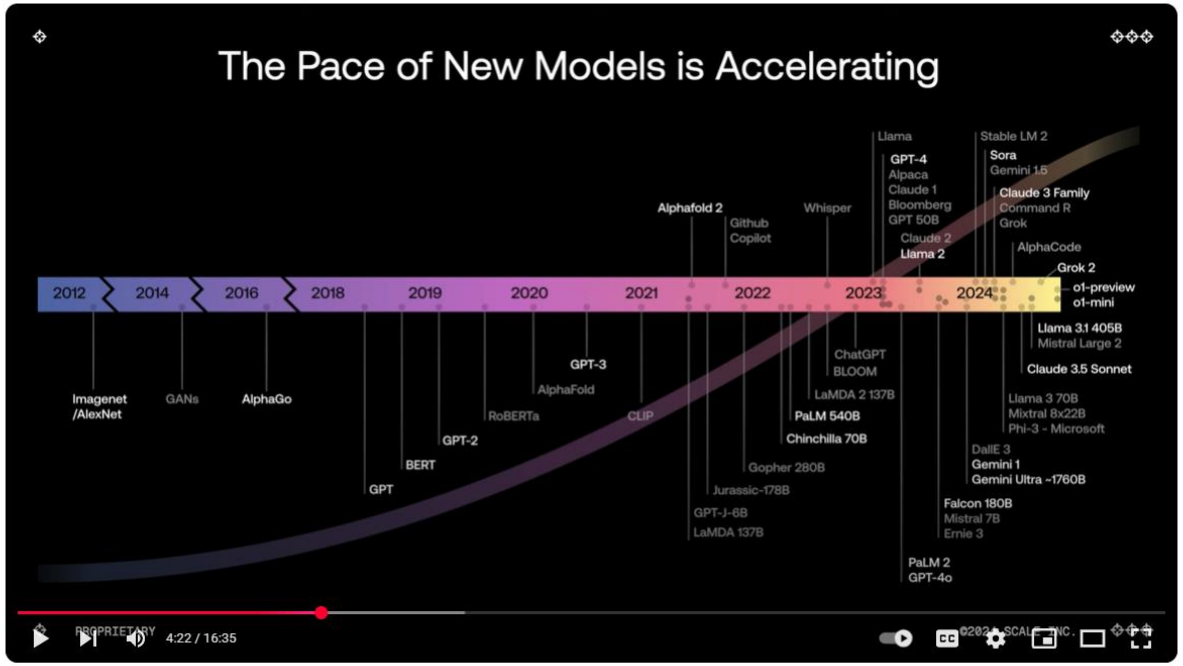
LLM Models

Let us explore a few models to better understand the LLM ecosystem. Each model has its own strengths and weaknesses and is designed for specific use cases.

Cohere   Meta   Mistral AI

AI21 labs AI21 Labs   a Amazon   A\ Anthrop

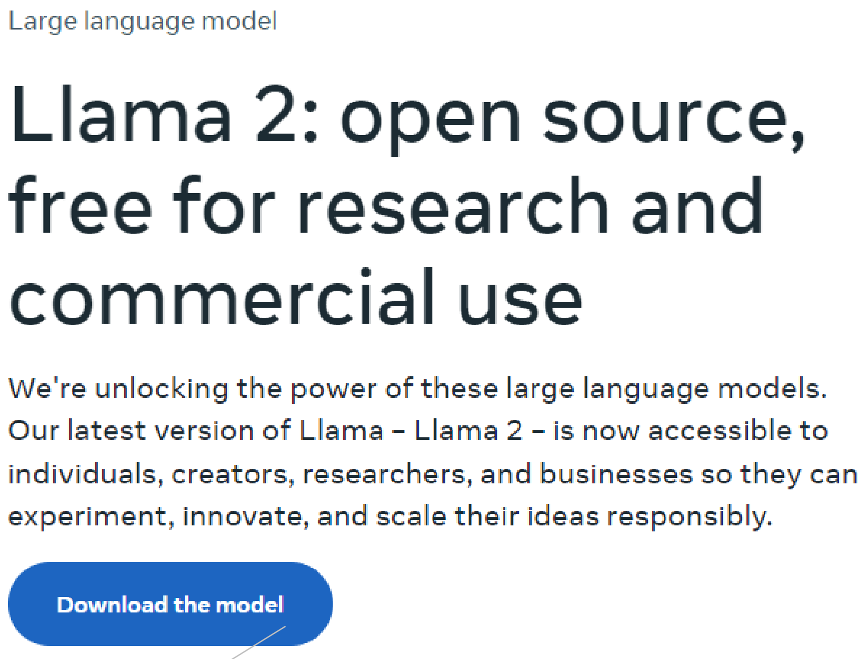
S. Stability AI   Gemini   OpenAI



The Pace of New Models is Accelerating

Scale AI Leadership Summit 2024: Alexandr Wang Opening Keynote

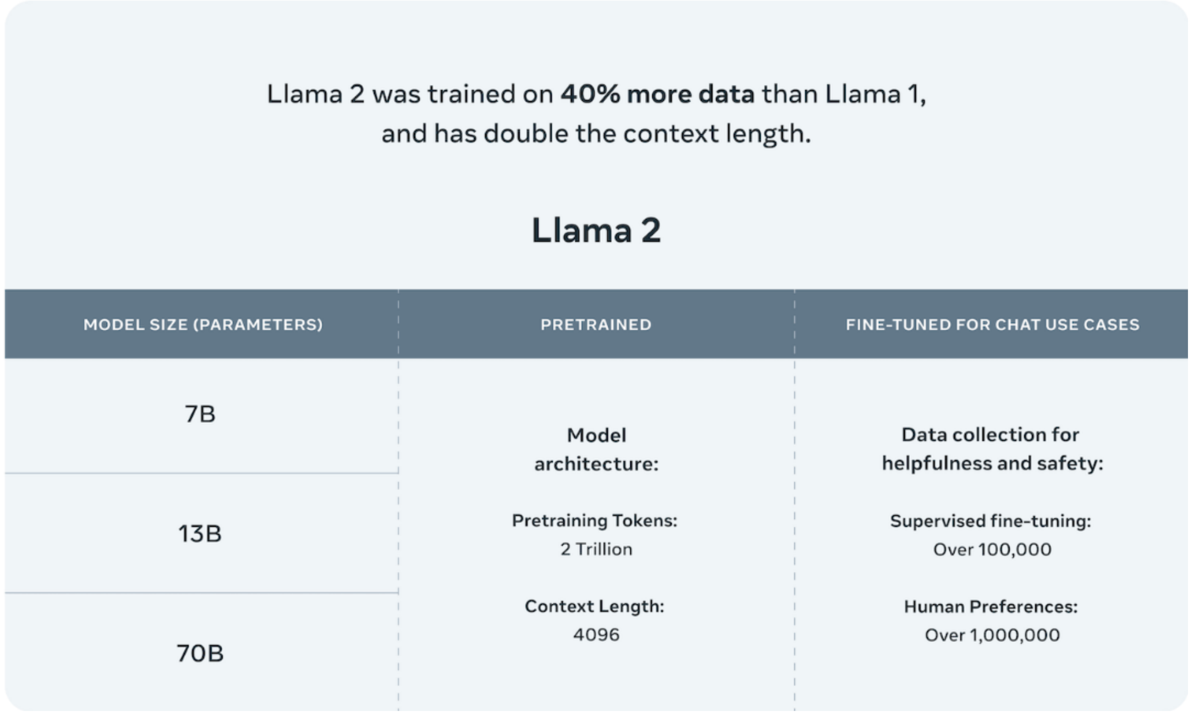
LLM models – Size, Context window, Parameters



Large language model

Llama 2: open source, free for research and commercial use

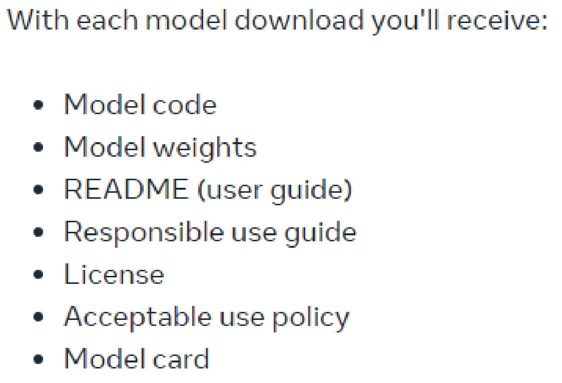
We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.



Llama 2 was trained on **40% more data** than Llama 1, and has double the context length.

Llama 2

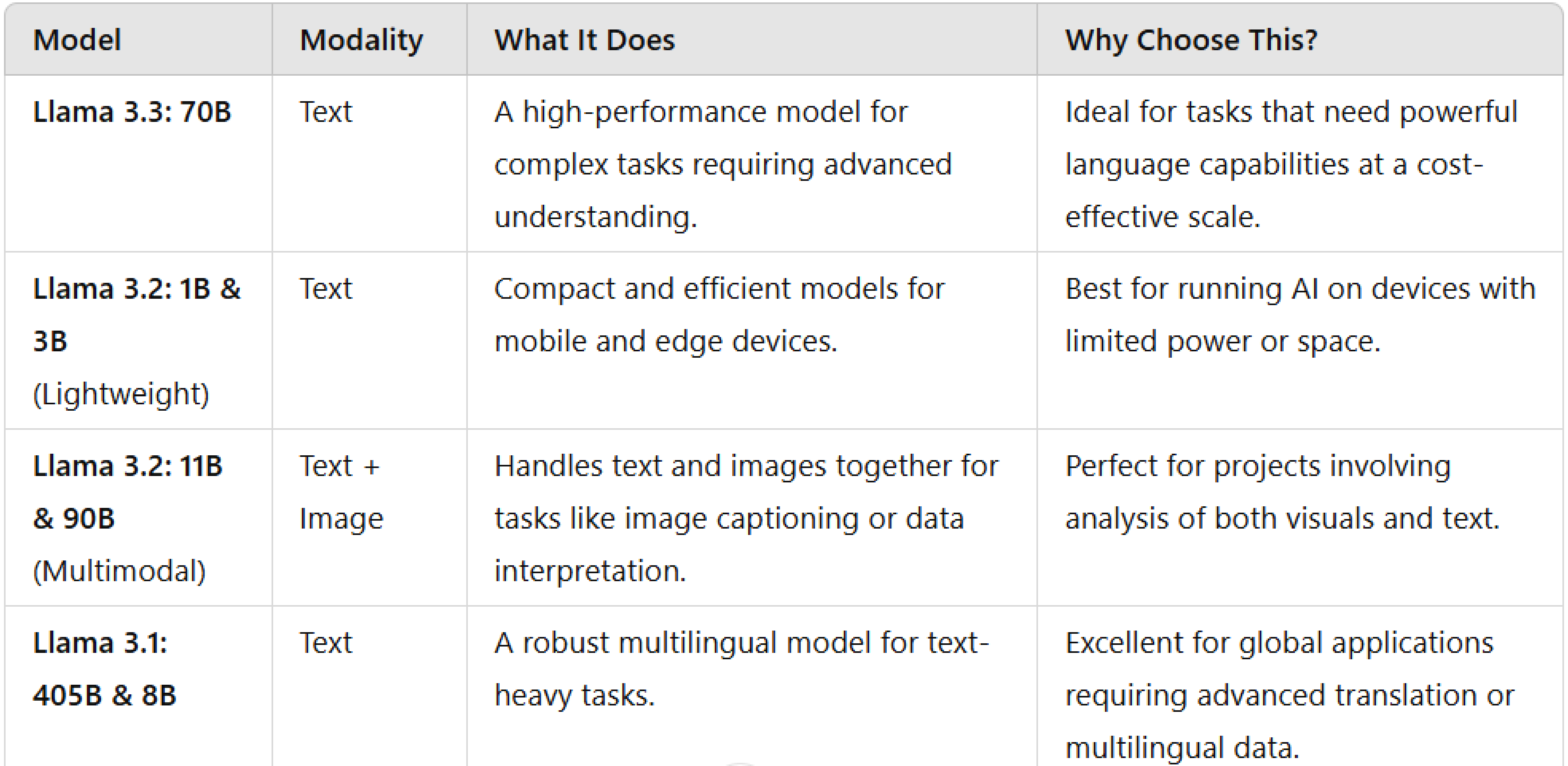
MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	**Model architecture:** **Pretraining Tokens:** 2 Trillion **Context Length:** 4096	**Data collection for helpfulness and safety:** **Supervised fine-tuning:** Over 100,000 **Human Preferences:** Over 1,000,000
13B		
70B		



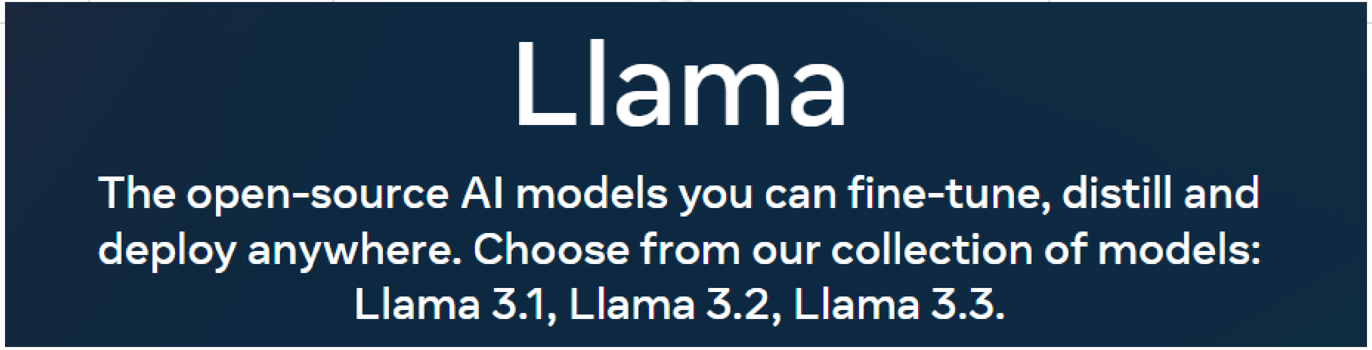
With each model download you'll receive:

- Model code
- Model weights
- README (user guide)
- Responsible use guide
- License
- Acceptable use policy
- Model card

Llama 3 open-source models from Meta



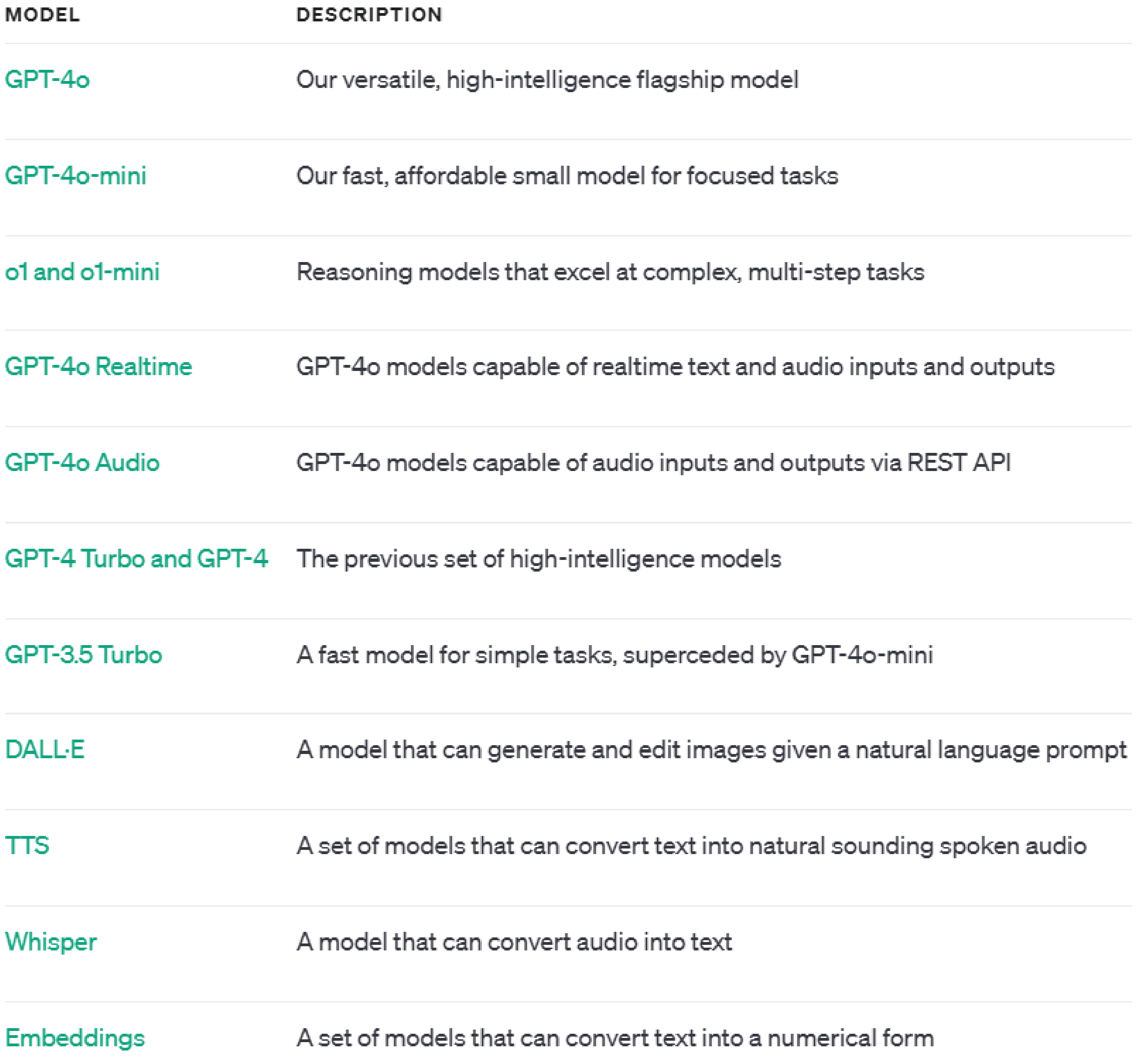
Model	Modality	What It Does	Why Choose This?
Llama 3.3: 70B	Text	A high-performance model for complex tasks requiring advanced understanding.	Ideal for tasks that need powerful language capabilities at a cost-effective scale.
Llama 3.2: 1B & 3B (Lightweight)	Text	Compact and efficient models for mobile and edge devices.	Best for running AI on devices with limited power or space.
Llama 3.2: 11B & 90B (Multimodal)	Text + Image	Handles text and images together for tasks like image captioning or data interpretation.	Perfect for projects involving analysis of both visuals and text.
Llama 3.1: 405B & 8B	Text	A robust multilingual model for text-heavy tasks.	Excellent for global applications requiring advanced translation or multilingual data.



Llama

The open-source AI models you can fine-tune, distill and deploy anywhere. Choose from our collection of models: Llama 3.1, Llama 3.2, Llama 3.3.

OpenAI models



MODEL	DESCRIPTION
GPT-4o	Our versatile, high-intelligence flagship model
GPT-4o-mini	Our fast, affordable small model for focused tasks
o1 and o1-mini	Reasoning models that excel at complex, multi-step tasks
GPT-4o Realtime	GPT-4o models capable of realtime text and audio inputs and outputs
GPT-4o Audio	GPT-4o models capable of audio inputs and outputs via REST API
GPT-4 Turbo and GPT-4	The previous set of high-intelligence models
GPT-3.5 Turbo	A fast model for simple tasks, superceded by GPT-4o-mini
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form

Source: https://platform.openai.com/docs/models

What makes LLMs special

Large language models like GPT-4 or Llama 3 have state-of-the-art capabilities such as general **knowledge**, **steerability**, **advanced reasoning**, **math/science**, **tool use**, **data analysis**, **multilingual translation** and more.

Based on transformer architecture LLM models are giants and can learn to understand human knowledge without supervision & without labelled datasets.

A single LLM model can perform multiple tasks such as QA, summarization, content/code generation, data analysis, translation and more

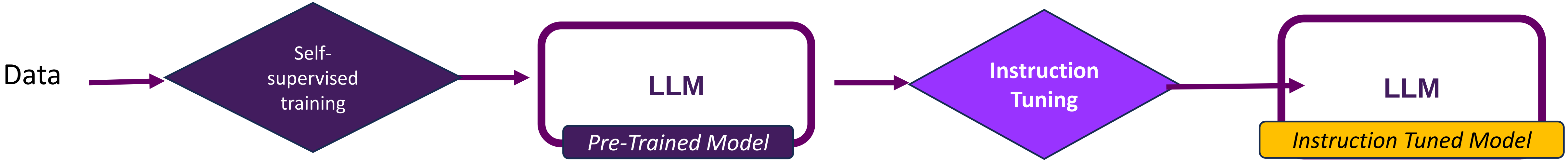
Models can be tuned to perform tasks for which they were never trained on.

LLM models can learn/understand patterns and representation of any sequence be it language, protein, biology, chemistry, etc.

LLMs are excellent few-shot learners. Using prompt engineering you can guide them to your request. LLMs can be multi-modal and so can be used in endless possible applications

How are LLMs trained?

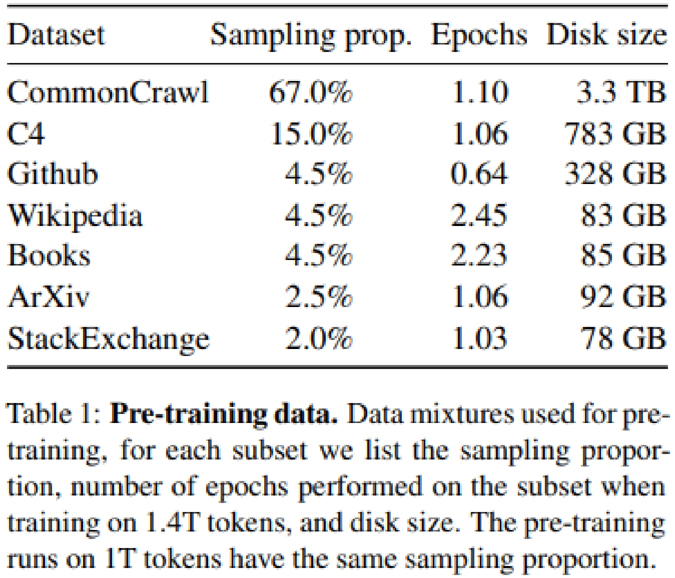
LLMs are very large deep learning models trained on huge amount of data. LLMs have a broad understanding of language, context, and world knowledge.



Both pre-trained and instruction-tuned models are foundation models. Because they are both built on a broad base of knowledge and are adaptable to a wide range of applications. The main difference is in the additional layer of training for instruction-tuned models, which is designed to enhance their ability to follow explicit instructions and perform tasks across different domains.

LLM Training : Stage 1 (Pretraining)

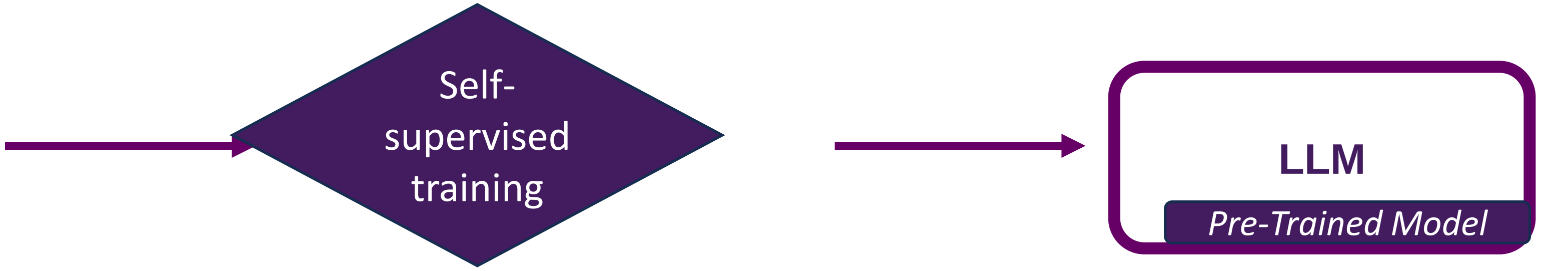
LLMs are trained on massive corpus of internet data using GPUs



Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Source (Paper on arxiv.org):
LLaMA: Open and Efficient Foundation Language Models



Self-supervised training

LLM

Pre-Trained Model

By processing vast amounts of text data, the model learns language, patterns, relationships between words, grammar, syntax, and aspects of knowledge contained within the text.

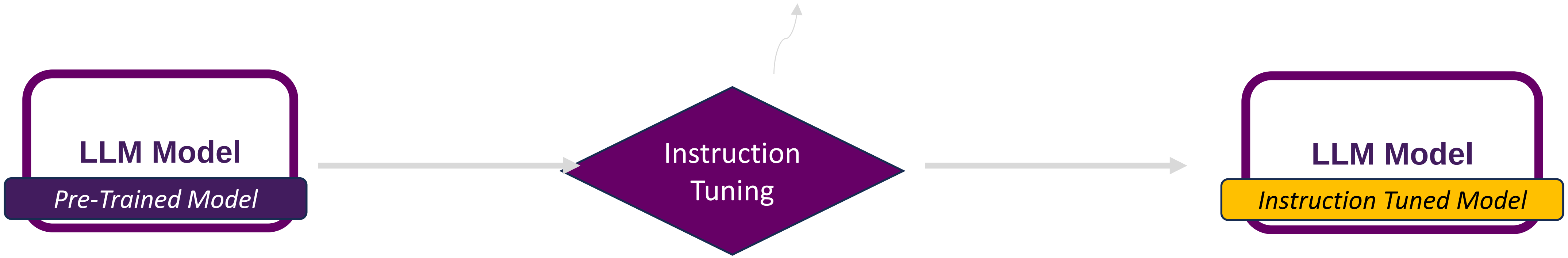
Have a broad understanding of language, context, and world knowledge. This general capability makes pretrained models versatile and adaptable to a variety of tasks with further fine-tuning or specific instructions. The model weights have been adjusted

LLM Training : Stage 2 (Post-training)

Supervised fine-tuning: Model is trained on low quantity/high quality labelled data such as Ideal Question/Response with human assistance

⊕

RLHF (Reinforcement Learning from Human Feedback)
Humans rank different responses generated by the model. This rating is captured in another model called the “reward model”. Then the LLM model is trained with the help of the “reward model” to generate responses



LLM Model

Pre-Trained Model

Instruction Tuning

LLM Model

Instruction Tuned Model

Can respond in a question & answer format. They excel at understanding and executing tasks based on natural language instructions.

Foundation Models – Summary: let us get comfortable

LLM Model

Pre-Trained Model

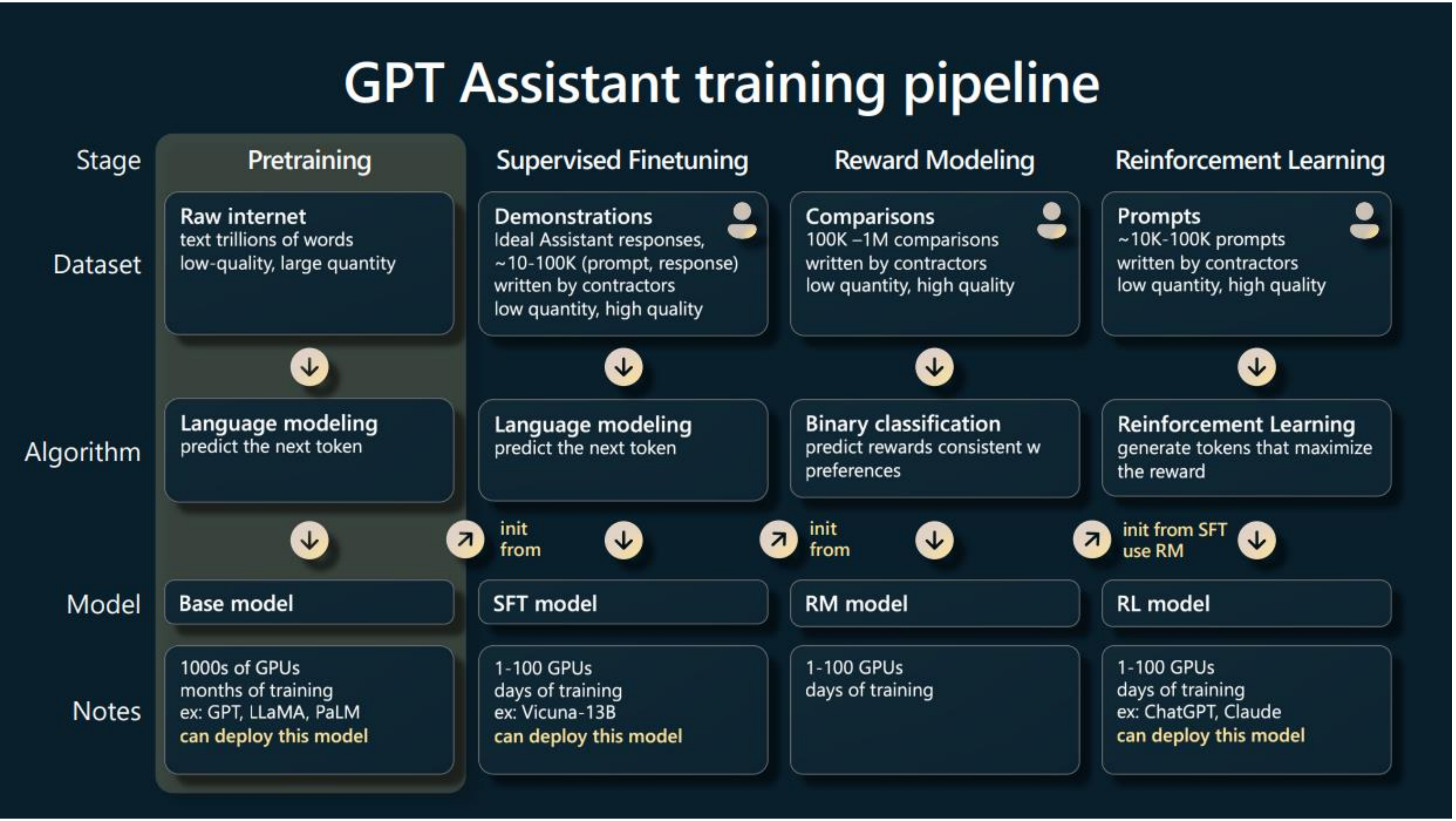
- **Definition:** These are models that have been initially trained on a large dataset to learn a wide range of patterns, knowledge, and language from that data. The process usually involves unsupervised learning, where the model learns to predict parts of the input (like the next word in a sentence) without explicit human-labeled instructions.
- **Purpose:** The main aim is to capture a broad understanding of language, context, and world knowledge. This general capability makes pretrained models versatile and adaptable to a variety of tasks with further fine-tuning or specific instructions.

LLM Model

Instruction Tuned Model

- **Definition:** These models start as pretrained models but undergo an additional phase of training (called instruction tuning or instruct-tuning) where they learn to follow human-like instructions or prompts more effectively. This stage involves supervised learning, typically using datasets where inputs are paired with instructions and desired outputs.
- **Purpose:** The goal is to improve the model's ability to understand and execute complex instructions given in natural language, making it more user-friendly and effective for tasks specified by users through prompts.

How was GPT trained ?



GPT Assistant training pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	**Base model**	**SFT model**	**RM model**	**RL model**
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM **can deploy this model**	1-100 GPUs days of training ex: Vicuna-13B **can deploy this model**	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude **can deploy this model**

State of GPT | BRK216HFS

688K views • 1 year ago

Must watch this video by

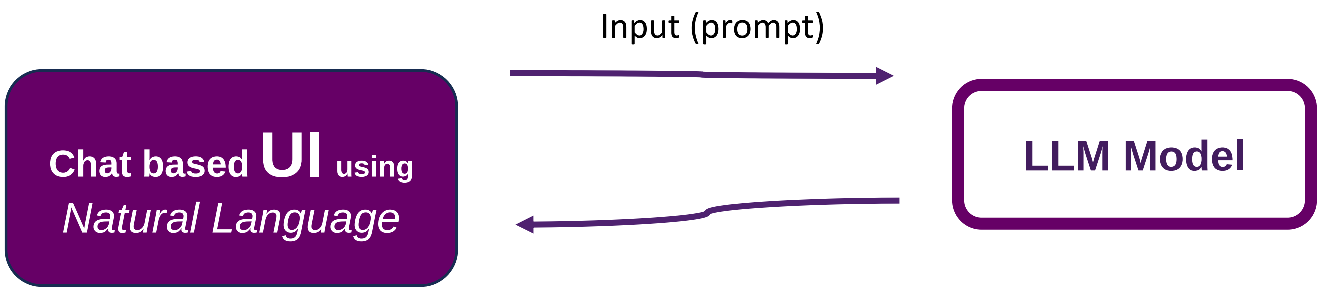
Andrej Karpathy .

He worked at OpenAI and Tesla.

. https://www.youtube.com/watch?v=bZQun8Y4L2A

AI systems with LLMs

LLMs have shown great promise as capable AI assistants for humans. LLMs can create new content, including text, images, videos, that can resemble works made by humans. These AI systems will be widely used for creativity, productivity, automation, and augmenting human work. In the next few years, the entire tech stack will be refactored. And as a result, the way we work will change.



- ChatGPT
- CoPilot App

Broad use cases

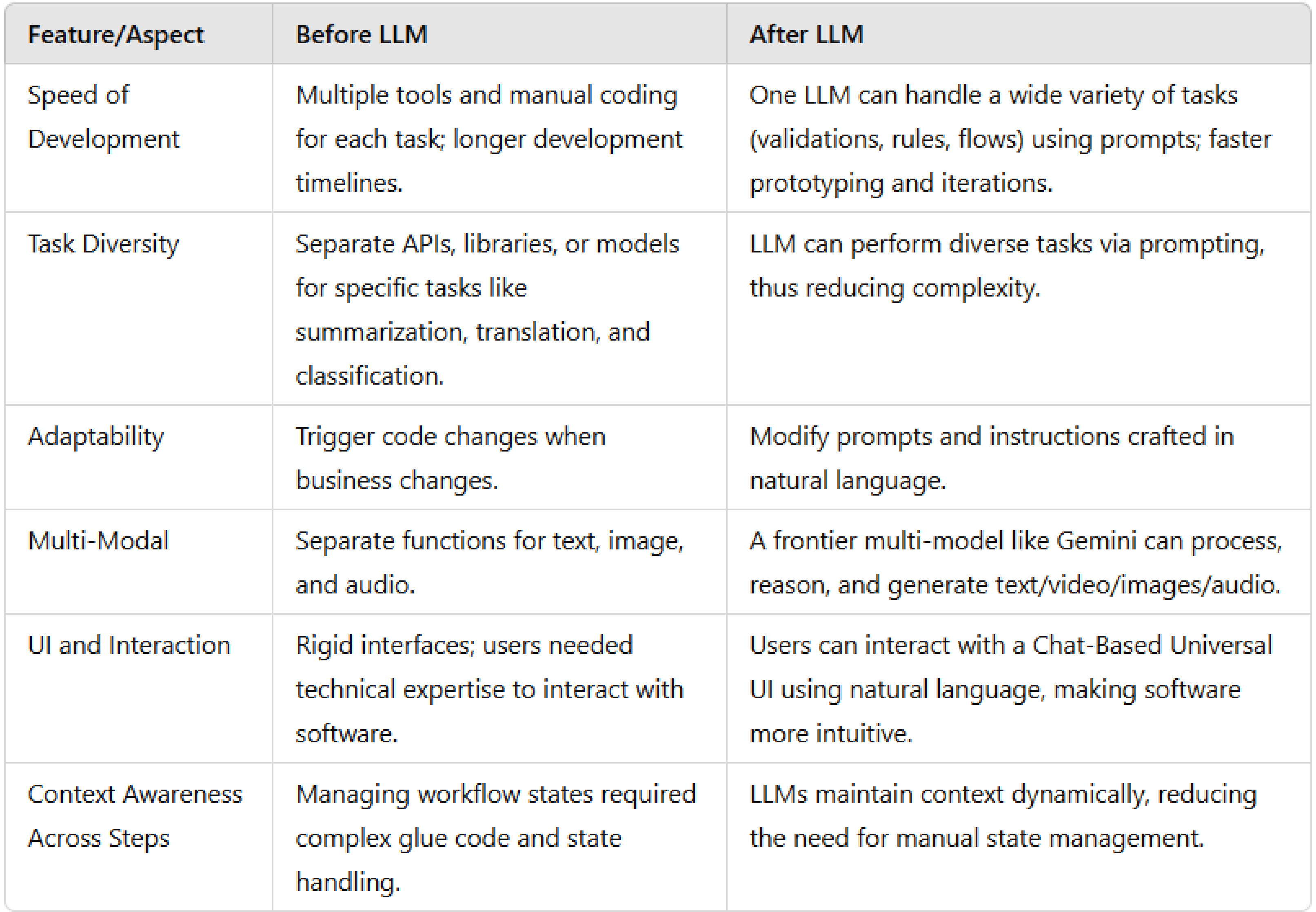
- Document summarization
- Coding assistant
- Extract knowledge via Q&A
- Personal tutor (for math's, physics, programming, etc.)
- Virtual assistant
- Idea generation
- Language translation
- Data analysis
- Agents
- Workflow automation
- Content generation (creative writing, images, videos, code)

and more

GPT-4 Large Claude 2 Gemini Pro GPT-3.5 LLaMA 2 70B

LLMs are instructible universal functions

In the next few years, the entire tech stack will be refactored. This will change the we work.



Feature/Aspect	Before LLM	After LLM
Speed of Development	Multiple tools and manual coding for each task; longer development timelines.	One LLM can handle a wide variety of tasks (validations, rules, flows) using prompts; faster prototyping and iterations.
Task Diversity	Separate APIs, libraries, or models for specific tasks like summarization, translation, and classification.	LLM can perform diverse tasks via prompting, thus reducing complexity.
Adaptability	Trigger code changes when business changes.	Modify prompts and instructions crafted in natural language.
Multi-Modal	Separate functions for text, image, and audio.	A frontier multi-model like Gemini can process, reason, and generate text/video/images/audio.
UI and Interaction	Rigid interfaces; users needed technical expertise to interact with software.	Users can interact with a Chat-Based Universal UI using natural language, making software more intuitive.
Context Awareness Across Steps	Managing workflow states required complex glue code and state handling.	LLMs maintain context dynamically, reducing the need for manual state management.

LLMs have made technology easier and faster to **develop** and **use**. Before LLMs, tasks required many tools and technical skills. Now, LLMs handle multiple tasks with simple instructions, making software use more intuitive and efficient.

LLMs can be instructed to perform tasks for which they were never trained on. LLMs are excellent few-shot learners. Using prompt engineering you can guide & steer them to fulfil your request in real-time.

LLM inferencing

The process of invoking LLMs in applications is called **inferencing**.

LLMs can be used in apps via:

- **API**: Connect to LLM services online for easy access.
- **On-Premise**: Deploy on local servers for more control and privacy.
- **Edge Computing**: Run on local devices for low latency and offline use.

Each method balances performance, cost, and privacy differently. Small size models are more suitable for edge inferencing.

Deployment Options: API, On-premise or On-device

Your business use case and choice of deployment will play a key role in model selection process.

On-Device

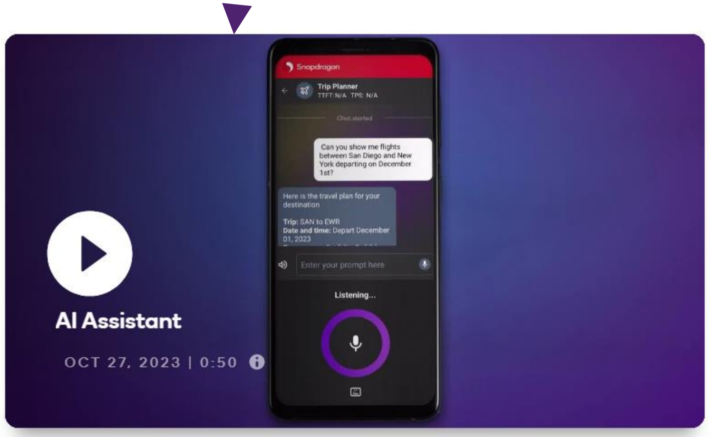
Smaller models for on-device processing like LLMA 2 7B.
Ex voice assistants

On-premise

Open-source models like LLMA 3 which you can download, modify & setup on a server in your company data center

Accessed via API

- Models like GPT-4 which can be accessed only via API calls.
- **On public cloud using AWS, Azure, GCP, etc.**

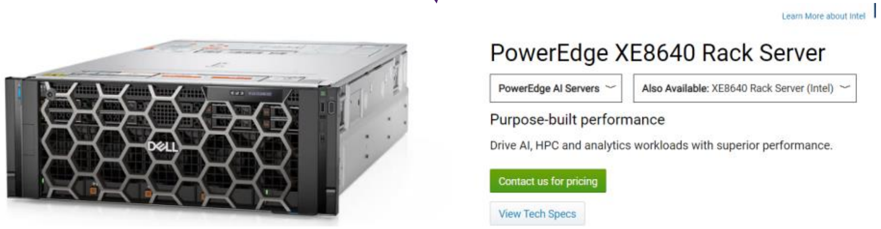


Can you show me flights between San Diego and New York departing on December 1st?

Here is the travel plan for your destination

Trip: SAN to EWR
Date and time: Depart December 01, 2023

Source:
www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai



PowerEdge XE8640 Rack Server

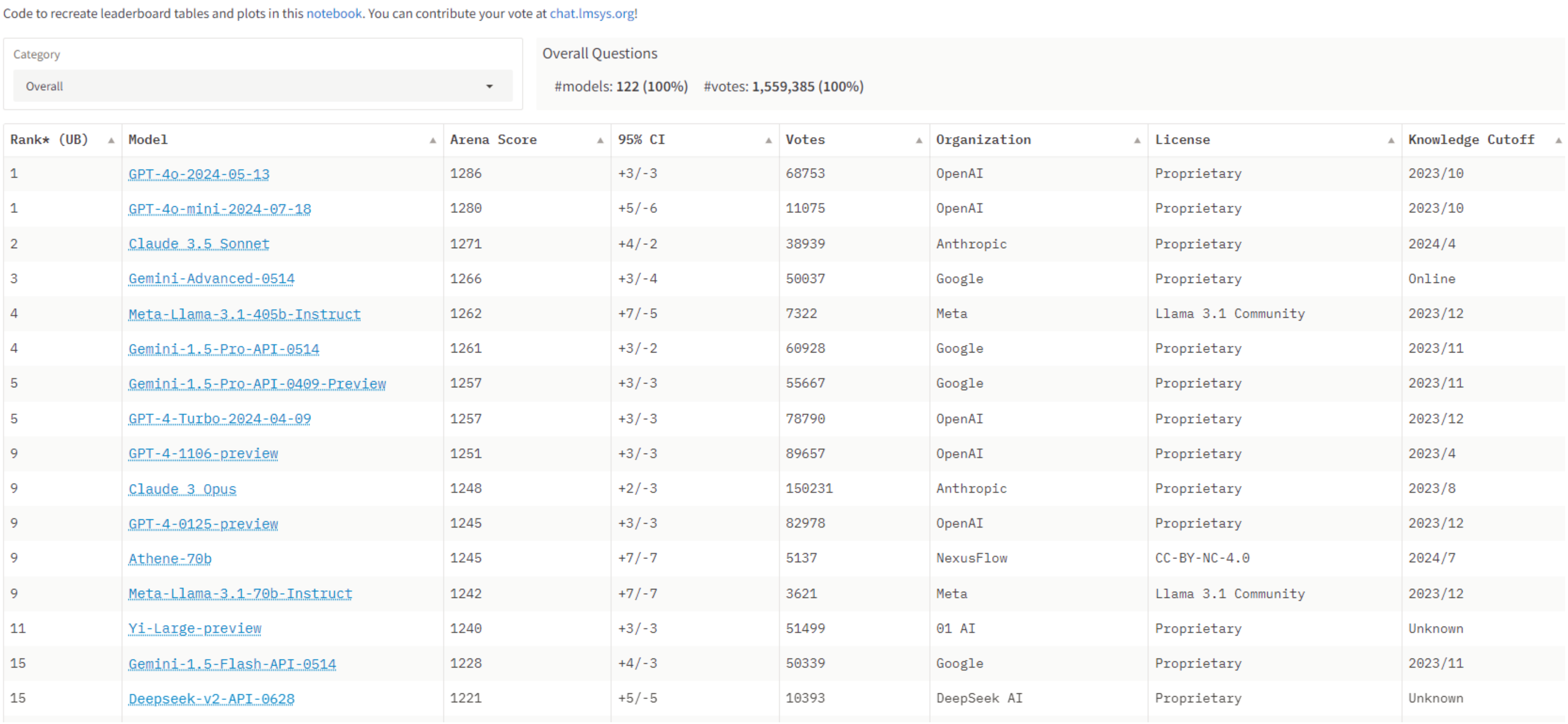
Purpose-built performance

Drive AI, HPC and analytics workloads with superior performance.

Source
www.dell.com/en-us/shop/ipovw/poweredge-xe8640?hve=shop+now

LLM Leaderboards

LLM leaderboards rank and compare language models based on performance metrics, track advancements, encourage innovation, and help users choose the best models for their needs.



Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

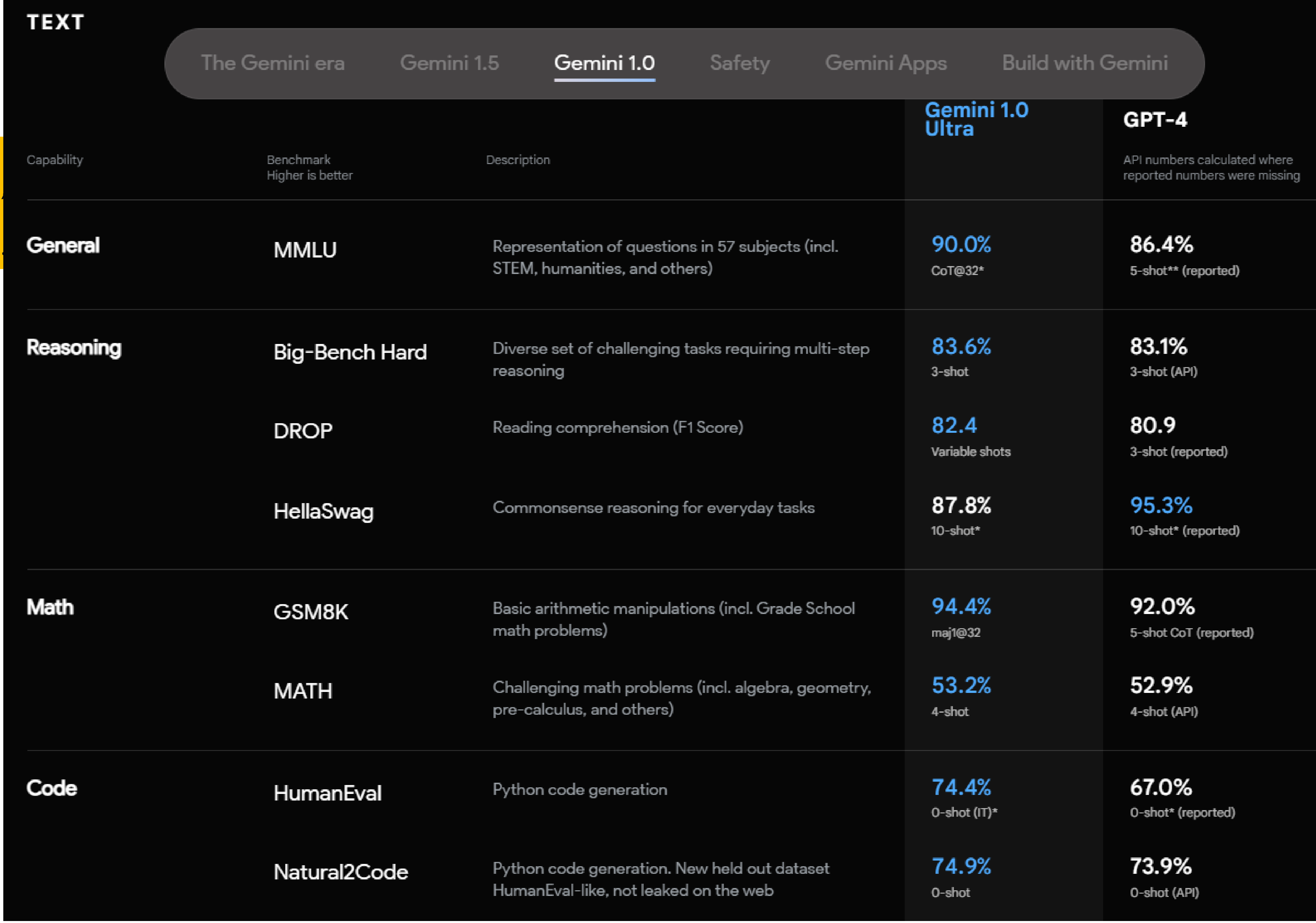
Overall Questions

#models: 122 (100%) #votes: 1,559,385 (100%)

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1286	+3/-3	68753	OpenAI	Proprietary	2023/10
1	GPT-4o-mini-2024-07-18	1280	+5/-6	11075	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1271	+4/-2	38939	Anthropic	Proprietary	2024/4
3	Gemini-Advanced-0514	1266	+3/-4	50037	Google	Proprietary	Online
4	Meta-Llama-3.1-405b-Instruct	1262	+7/-5	7322	Meta	Llama 3.1 Community	2023/12
4	Gemini-1.5-Pro-API-0514	1261	+3/-2	60928	Google	Proprietary	2023/11
5	Gemini-1.5-Pro-API-0409-Preview	1257	+3/-3	55667	Google	Proprietary	2023/11
5	GPT-4-Turbo-2024-04-09	1257	+3/-3	78790	OpenAI	Proprietary	2023/12
9	GPT-4-1106-preview	1251	+3/-3	89657	OpenAI	Proprietary	2023/4
9	Claude 3 Opus	1248	+2/-3	150231	Anthropic	Proprietary	2023/8
9	GPT-4-0125-preview	1245	+3/-3	82978	OpenAI	Proprietary	2023/12
9	Athene-70b	1245	+7/-7	5137	NexusFlow	CC-BY-NC-4.0	2024/7
9	Meta-Llama-3.1-70b-Instruct	1242	+7/-7	3621	Meta	Llama 3.1 Community	2023/12
11	Yi-Large-preview	1240	+3/-3	51499	01 AI	Proprietary	Unknown
15	Gemini-1.5-Flash-API-0514	1228	+4/-3	50339	Google	Proprietary	2023/11
15	Deepseek-v2-API-0628	1221	+5/-5	10393	DeepSeek AI	Proprietary	Unknown

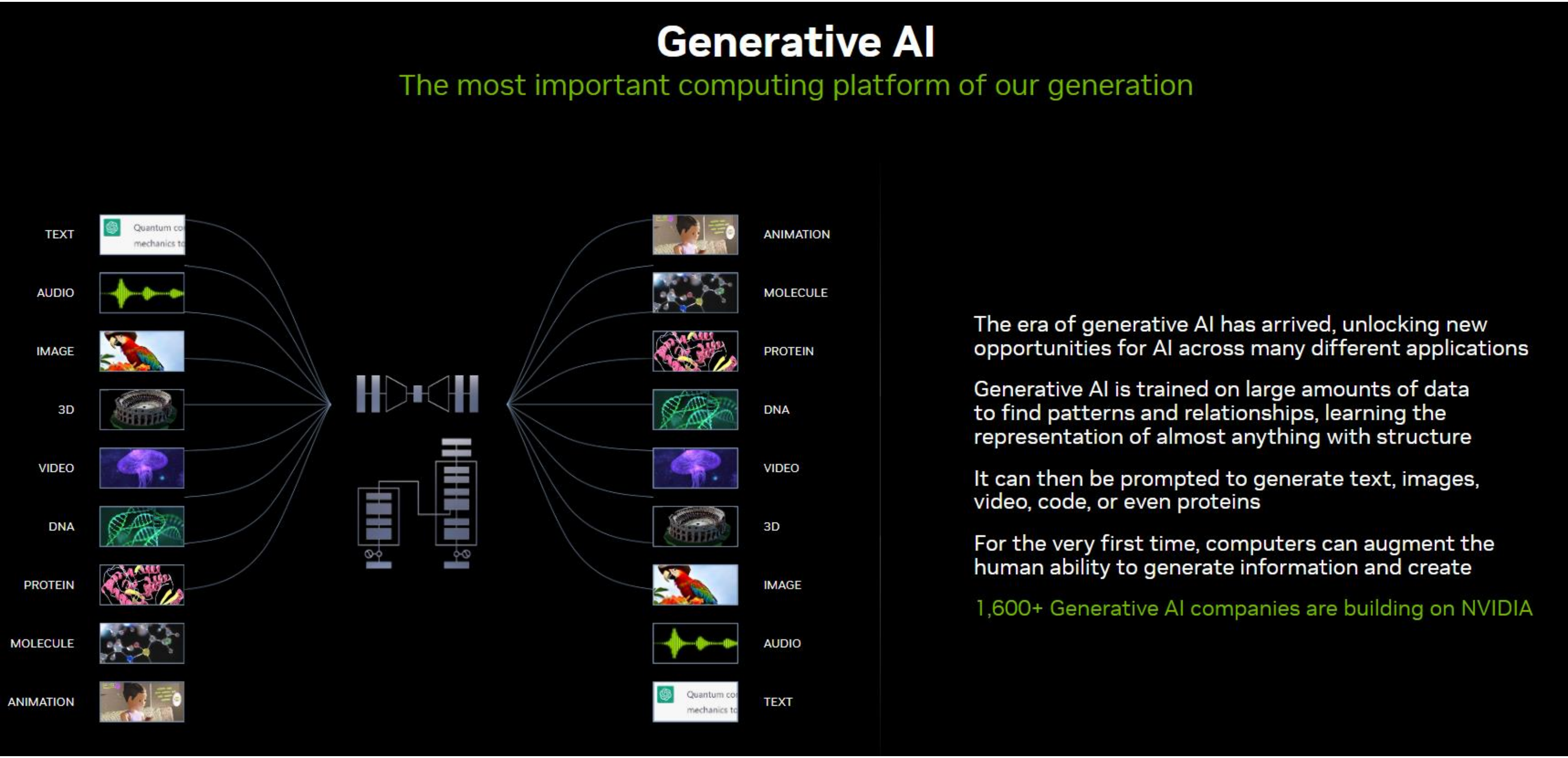
https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

LLM benchmarking refers to the evaluation of large language models, to assess their performance & efficiency across a variety of metrics.



Capability	Benchmark Higher is better	Description	Gemini 1.0 Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	95.3% 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	94.4% maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	53.2% 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)

Source: https://deepmind.google/technologies/gemini/#gemini-1.0



Generative AI

The most important computing platform of our generation

The era of generative AI has arrived, unlocking new opportunities for AI across many different applications

Generative AI is trained on large amounts of data to find patterns and relationships, learning the representation of almost anything with structure

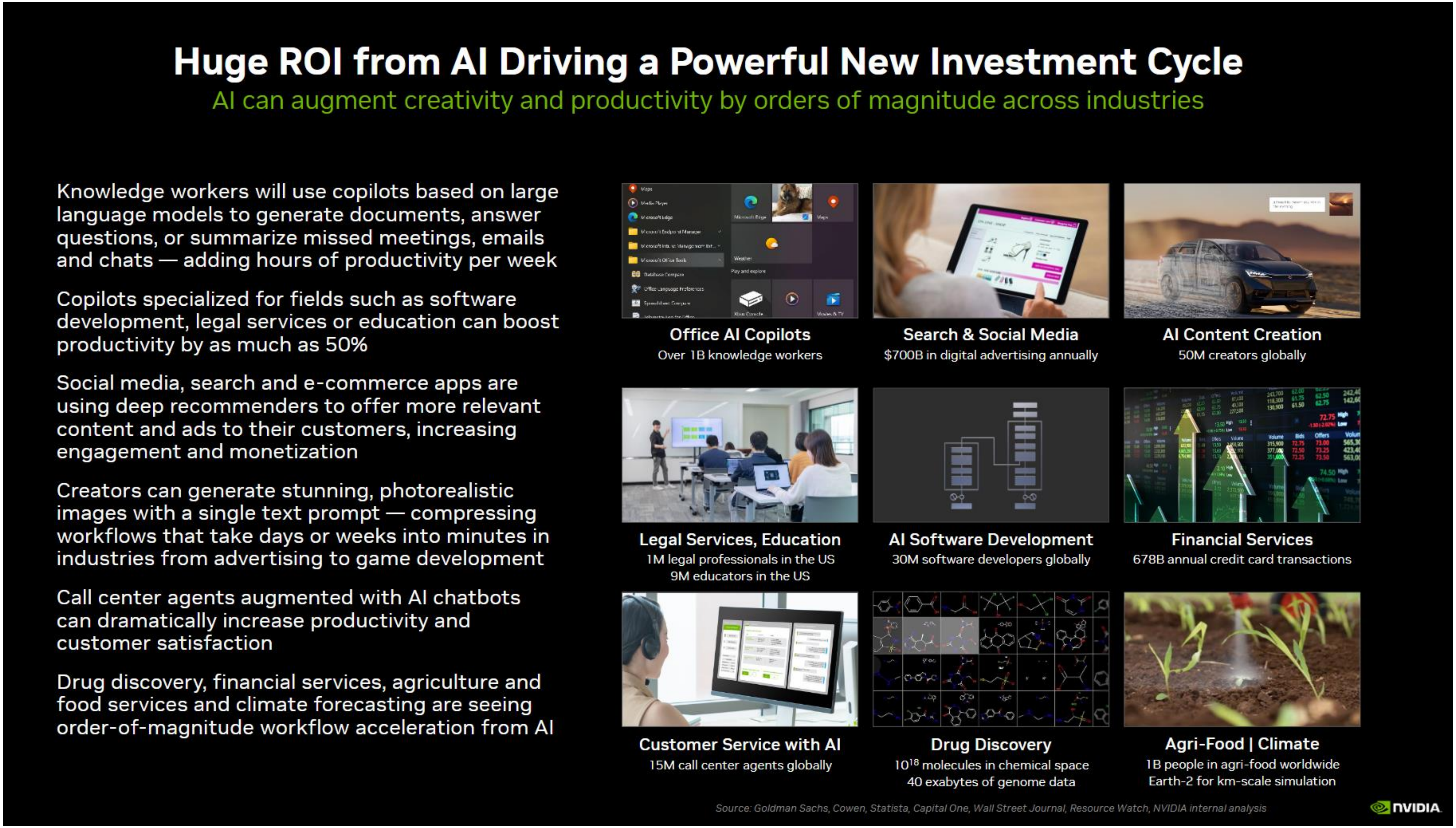
It can then be prompted to generate text, images, video, code, or even proteins

For the very first time, computers can augment the human ability to generate information and create

1,600+ Generative AI companies are building on NVIDIA

Source: Nvidia

Importance of Generative AI



Huge ROI from AI Driving a Powerful New Investment Cycle

AI can augment creativity and productivity by orders of magnitude across industries

Knowledge workers will use copilots based on large language models to generate documents, answer questions, or summarize missed meetings, emails and chats — adding hours of productivity per week

Copilots specialized for fields such as software development, legal services or education can boost productivity by as much as 50%

Social media, search and e-commerce apps are using deep recommenders to offer more relevant content and ads to their customers, increasing engagement and monetization

Creators can generate stunning, photorealistic images with a single text prompt — compressing workflows that take days or weeks into minutes in industries from advertising to game development

Call center agents augmented with AI chatbots can dramatically increase productivity and customer satisfaction

Drug discovery, financial services, agriculture and food services and climate forecasting are seeing order-of-magnitude workflow acceleration from AI

Office AI Copilots
Over 1B knowledge workers

Search & Social Media
$700B in digital advertising annually

AI Content Creation
50M creators globally

Legal Services, Education
1M legal professionals in the US
9M educators in the US

AI Software Development
30M software developers globally

Financial Services
678B annual credit card transactions

Customer Service with AI
15M call center agents globally

Drug Discovery
10^{18} molecules in chemical space
40 exabytes of genome data

Agri-Food | Climate
1B people in agri-food worldwide
Earth-2 for km-scale simulation

Source: Goldman Sachs, Cowen, Statista, Capital One, Wall Street Journal, Resource Watch, NVIDIA internal analysis

Source: blogs.nvidia.com/blog/llms-ai-horizon

Importance of Gen AI

Importance of **Generative AI**

Improve productivity
Eliminate drudgery
Your reasoning engine
Increase innovation
Transform business
Personal Assistant

CoPilots & Assistants

Empower humans in their line of work in business. Personal tutor.

Universal UI

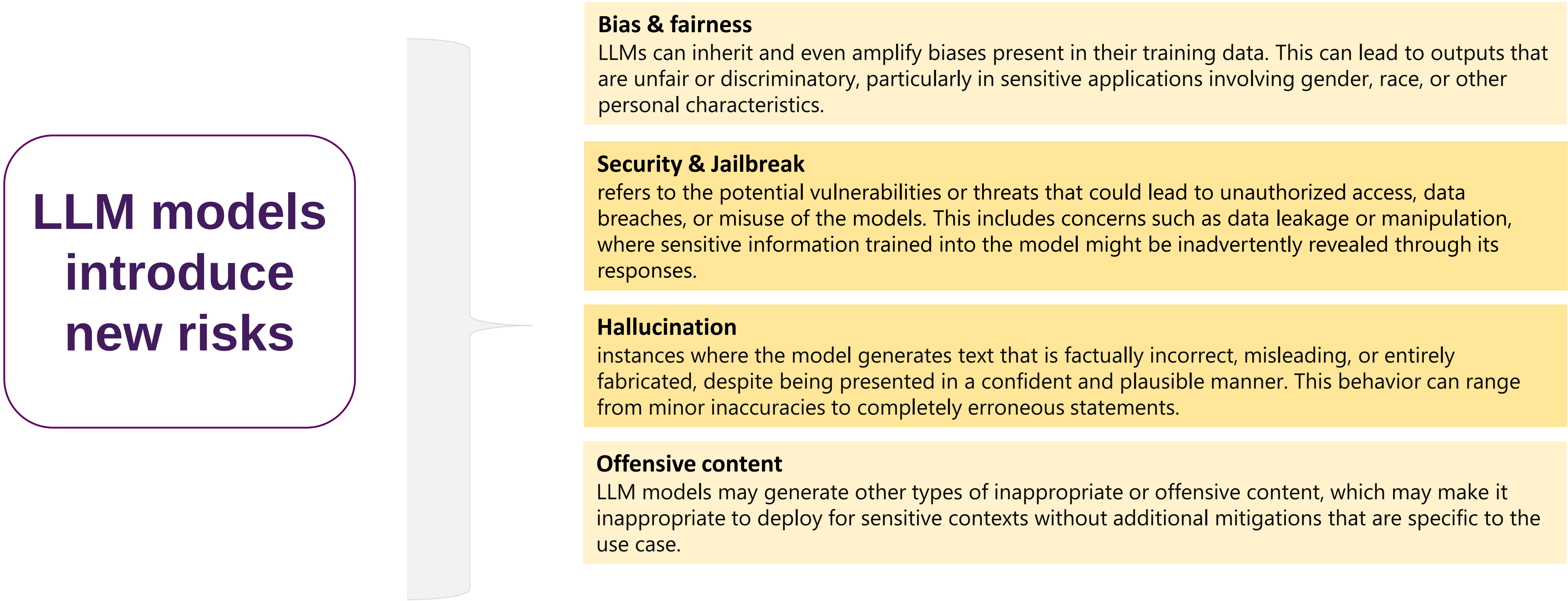
Natural language is the new interface for text, speech or video. Humans will learn & cocreate with AI using natural language

AI Orchestrator: AI Agents

LLMs can function as AI orchestrators by coordinating the interaction between various systems & services.

Gen AI introduces new risks

Gen AI offer great promise but comes with risks related to responsible AI. Gen AI systems can cause harm such as promote misinformation, hallucinate, etc. and lead to a wide range of other negative impacts..



LLM models introduce new risks

Bias & fairness
LLMs can inherit and even amplify biases present in their training data. This can lead to outputs that are unfair or discriminatory, particularly in sensitive applications involving gender, race, or other personal characteristics.

Security & Jailbreak
refers to the potential vulnerabilities or threats that could lead to unauthorized access, data breaches, or misuse of the models. This includes concerns such as data leakage or manipulation, where sensitive information trained into the model might be inadvertently revealed through its responses.

Hallucination
instances where the model generates text that is factually incorrect, misleading, or entirely fabricated, despite being presented in a confident and plausible manner. This behavior can range from minor inaccuracies to completely erroneous statements.

Offensive content
LLM models may generate other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.