The Generative AI Revolution: Exploring the Current Landscape

Towards AI Editorial Team · Follow
Published in Towards AI
30 min read · 4 days ago

Listen Share

Generative AI has gained extensive attention and investment in the past year due to its ability to produce coherent text, images, code, and beyond-impressive outputs with just a simple textual prompt. However, the potential of this generation of AI models goes beyond typical natural language processing (NLP) tasks. There are countless use cases, such as explaining complex algorithms, building bots, helping with app development, and explaining academic concepts. Fields like animation, gaming, art, movies, and architecture are being revolutionized by text-to-image programs like DALL-E, Stable Diffusion, and Midjourney. Additionally, generative AI models have shown transformative capabilities in complex fields like software development, with tools such as GitHub Copilot and Replit Ghostwriter.

While today's generative models are built upon a decade of progress, 2022 was the year when generative AI triggered an "Aha!" moment. Based on discussions about this new era of human-machine cooperation, important questions arise, such as *why now, and what's next?* This post explains the journey of how it all started, where it is going, and some of the biggest players and most popular models in the Generative AI landscape today, along with real-world tools designed for users to optimize creation, ideation, development, and production processes.

What is Generative AI?

Generative AI is a subfield of machine learning that involves training artificial intelligence models on large volumes of real-world data to generate new contents (text, image, code,...) that is comparable to what humans would create. This is achieved by training algorithms on large datasets to identify patterns and learn from them. Once the neural network has learned these patterns, it can generate new data that adheres to the same patterns. However, this process is computationally intensive.

Fundamentally, a generative AI for NLP applications will process an enormous corpus on which it has been trained and respond to prompts with something that falls within the realm of probability, as learnt from the mentioned corpus. For example, autocomplete is a low-level form of generative AI. Advanced models like ChatGPT and DALL-E take the concept to a whole new level. Different model architectures, such as diffusion models and Transformer-based large language models (LLMs), can be employed for generative tasks such as image and language generation.

Diffusion models are a type of generative AI model that can be used for a variety of tasks, including image generation, image denoising, and inpainting. Similarly, the Transformer architecture revolutionized the language domain. The new era of language models are Transformer-based, which is a type of deep learning architecture for natural language processing (NLP) tasks. They utilize a self-attention mechanism to transform the input sequence into a set of context-aware high dimensional vectors (also known as embeddings) that can be used for a variety of NLP tasks, including language generation, machine translation, and text classification. The most well-known transformer-based LLMs are the GPT family, developed by OpenAI. The primary advantage of transformer-based LLMs over traditional NLP models is that they are highly parallelizable and can handle long-range dependencies between words in a sentence more effectively. This makes them more suitable for tasks that require a deeper understanding of the context, such as text summarization or generating a coherent and fluent text.

Let's explore the history and current state of generative AI and the key players shaping its future.

The Generative AI Revolution

Generative AI has been around for several years. One of the earliest examples is the Eliza chatbot developed by Joseph Weizenbaum in 1966. However, these early implementations relied on a rules-based approach that had several shortcomings, such as a limited vocabulary, lack of context, and overreliance on patterns. As a result, they were prone to frequent breakdowns, making it difficult to customize and expand these initial chatbots.

Recently, significant progress has been made in AI and machine learning, resulting in the development of advanced generative AI systems. It's no coincidence that these breakthroughs have happened all at once. They're based on a new class of AI models that are incredibly flexible and powerful, surpassing anything we've seen before. In deep learning, there are

three critical components that contributed the most to their recent success: scaling models, large datasets, and more compute power — all working together to bring us to this exciting point in AI advancement.

Progress in GPUs and their application to Machine Learning

GPUs are designed for parallel processing, making them well-suited for the computationally intensive tasks involved in training deep neural networks. Unlike CPUs, which focus on sequential processing, GPUs have thousands of smaller cores that can handle multiple tasks simultaneously, allowing for faster training of large networks. A key breakthrough for machine learning was the intuition that GPUs could be used for Neural Networks, together with software progress such as Nvidia's release of CUDA in 2007, a programming language that allowed GPUs to be used as general-purpose computers.

Alexnet — 2012 — The Deep Learning Revolution

The modern AI revolution began in 2012 with step change progress in deep learning and convolutional neural networks (CNNs), which were particularly effective in solving computer vision problems. Although CNNs had been around since the 1990s, they were not practical due to their intensive computing power requirements. However, In 2009, Stanford AI researchers introduced ImageNet, a labeled image dataset used to train computer vision algorithms, and a yearly challenge. In 2012, AlexNet combined CNNs trained on GPUs with ImageNet data to create the most advanced visual classifier at the time. The model outperformed the runner-up by a significant margin of nearly 11%! The success of CNNs, the ImageNet dataset, and GPUs drove significant progress in computer vision.

Transformers: Attention Is All You Need (Google) — 2017

One critical area where deep learning lagged was natural language processing (NLP), which involves getting computers to understand and hold a coherent conversation with humans rather than translation or classification. NLP breakthroughs were needed to bridge this gap. Previously, researchers relied on models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) to process and analyze time-based data. These models were proficient at recognizing short sequences such as spoken words but struggled with longer sentences and paragraphs. The architectural flaws of these models was unable to capture the complexity and richness of ideas that arise when sentences are combined into larger bodies of text.

A significant breakthrough in AI was the development of the "Transformer" model by Google with the very popular paper "Attention Is All You Need". This model represented a major milestone as it revolutionized the approach to translation problems by utilizing a mechanism called "attention": a particular neural network that allowed the model to analyze the entire input sequence and determine relevance to each component of the output. In the years to come, Transformers have been found to be state-of-the-art models for many other NLP tasks as well, and recently also in other domains such as computer vision.

Next word prediction, scale and fine tuning — BERT (Google) and GPT (OpenAI) family — 2018

With the advancement of Transformers, a key further breakthrough finding was the potential to train on unstructured data via next word prediction objective on website contents. It introduced the models such as BERT and GPT-2. This delivered surprising capabilities and "zero shot" performance at completing new tasks the model hadn't been trained for. OpenAI also continued to probe the ability for the performance of these models to continue increasing with more scale and more training data.

One of the major challenges faced by researchers was acquiring the right training data. ImageNet, a collection of one hundred thousand labeled images, required a significant human effort. Despite the abundance of text available on the Internet, creating a meaningful dataset for teaching computers to work with human language beyond individual words is a time-consuming process. Additionally, labels created for one application using the same data may not apply to another task. With the advancements of BERT and first iteration of GPT, we started to harness the immense amount of unstructured text data available on the internet and the computational power of GPUs. OpenAI further advanced this approach with their development of GPT-2 and GPT-3 models, which are short for "generative pre-trained transformer." These models are specifically designed to generate new words in response to input and are pre-trained on vast amounts of text using the next world prediction objective.

Another key breakthrough in these large transformer models is the concept of "fine tuning" — or adapting a large model to new more specific tasks or a new smaller and targeted data set — to improve performance in a particular domain with far lower compute cost than training a new model from scratch. For example, a foundational language model like GPT-3 may be fine-tuned on a dataset of medical documents to create an instruction-tuned model for medical document processing. This model will be better at understanding medical terminology, identifying medical entities, and extracting relevant information from medical texts.

Instruction Tuning — Instruct GPT and ChatGPT (OpenAI) — 2022

The most recent advancement which has led to the Generative AI landscape today is the concept of Instruction Tuning — taking a model which has just been trained to predict the next word of a text document — and teaching it (via fine tuning) to actually follow human instructions and preferences. This made it far easier to interact with these LLMs and to get them to answer questions and perform tasks without getting sidetracked by just trying to predict the next word. A fortunate feature of instruction tuning is that not only it helps to increase the accuracy and capabilities of these models, but they also help align them to human values and helps prevent them from generating undesired or dangerous content.

OpenAI's specific technique for instruction tuning is called reinforcement learning with human feedback (RLHF) where humans are used to train the model by ranking its responses. Building on top of Instruction Tuning, OpenAI released ChatGPT — which reorganized instruction tuning into a dialogue format and created an easy to use interface for interacting with the AIs. This has catalyzed the mass awareness and adoption of Generative AI products and has led to the landscape we have today.

The Current LLM Landscape

The breakthroughs in Generative AI have left us with an extremely active and dynamic landscape of players. This consists of 1) AI hardware manufacturers such as Nvidia and Google, 2) AI cloud platforms such as Azure, AWS, Nvidia and Google, 3) open source platforms for accessing the full models, such as Hugging Face, 4) access to LLM models via API such as OpenAI, Cohere and Anthropic and 5) access to LLMs via consumer products such as ChatGPT and Bing. Additionally, there are many more breakthroughs happening each week in this universe such as the release of multi modal models (that can understand both text and image), new model architectures (such as Mixture of Experts), Agent Models (models that can set tasks and interact with each other and other tolls).

This all leads to many questions such as;

- How will most people interact with LLMs?
- Who will be the leading players going forward?
- How fast will the capabilities of these models keep progressing?
- Are open source models dangerous because of the lack of control of their outputs and use, or are they beneficial due to democratizing access to this technology?

The Leading LLMs Models (broadly from low to high training cost)

Company	Model	Parameters (bn)	Training tokens (bn)	Release date
Meta	LLaMA	7	1000	February 2023
EleutherAI	NeoX	20	420	February 2022
Meta	Galactica	120	106	November 2022
Cohere	Cohere XLarge	52		February 2022
Anthropic	Anthropic-LM v4-s3	52		April 2022
Google	Google LaMDA	137	168	May 2021
Google	GLaM (Mixture of experts)	1200		December 2021
Google Deepmind	DeepMind Gopher	80	300	December 2021
Meta	OPT	175	180	May 2022
Open AI	GPT-3	175	300	June 2020
A121	A121 Jurassic-1	178	300	August 2021
BigScience	Bloom	176	366	August 2022
Baidu	Ernie 3.0 Titan	260		December 2021
Meta	LLaMA	63	1400	February 2023
Google Deepmind	DeepMind Chinchilla	70	1400	March 2022
Mosaic	MosaicML GPT	70	1400	September 2022
Nvidia & Microsoft	MT-NLG	530	270	October 2021
Google	PaLM	540	780	April 2022
Open AI	GPT-4			March 2023

1. OpenAI's GPT Models

Notable Models

Model	Function	Pricing	Link
GPT-4	More capable than any GPT family model, able to do more complex tasks, and optimized for chat.	*8K Context:* $0.03 / 1K tokens (prompt) $0.06 / 1K tokens (completion) *32K context:* $0.06 / 1K tokens (prompt) $0.12 / 1K tokens (completion)	https://platform.openai.com/docs/models/gpt-4
gpt-3.5-turbo	Optimized for dialogue, most capable GPT-3.5 model and optimized for chat at 1/10th the cost of text-davinci-003.	$0.002 / 1K tokens	https://platform.openai.com/docs/guides/chat

Task specific models

Model	Function	Training	Usage
Ada	Capable for simple tasks like classification (Fastest)	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Davinci	Most capable of other GPT-3 models (Most Powerful)	$0.0200 / 1K tokens	$0.1200 / 1K tokens
Babbage	Capable of straightforward tasks, fast, and lower cost.	$0.0005 / 1K tokens	$0.0024 / 1K tokens
Curie	Very capable, faster, and lower cost than Davinci.	$0.0020 / 1K tokens	$0.0120 / 1K tokens

Find model information here: https://platform.openai.com/docs/models/gpt-3

Image & Audio Models

Model	Function	Pricing	Link
DALL·E	Image model	$0.016 / image (256×256 resolution)	https://platform.openai.com/docs/guides/images
Whisper	Audio model	$0.006 / minute	https://platform.openai.com/docs/guides/speech-to-text

OpenAI, the company behind the GPT models, is an AI research and deployment company. The San Francisco-based lab was founded in 2015 as a nonprofit with the goal of building "artificial general intelligence" (AGI), which is essentially software as smart as humans. OpenAI conducts innovative research in various fields of AI, such as deep learning, natural language processing, computer vision, and robotics, and develops AI technologies and products intended to solve real-world problems.

OpenAI transitioned into a for-profit company in 2019. The company plans to cap the profit of the investors at a fixed multiple of their investment (noted by Sam Altman as currently ranging between 7x and 100x depending on the investment round date and risk). As per the WSJ OpenAI was initially funded by $130m of charity funding (Elon Musk tweeted he contributed $100m) and has since raised at least $13bn led by Microsoft (where OpenAI makes use of Azure cloud credits). With the Microsoft partnership, OpenAI's ChatGPT, along with Microsoft's own search AI, created an improved version of Bing and transformed Microsoft's Office productivity apps.

In 2019, OpenAI released GPT-2, a model that could generate realistic human-like text in entire paragraphs with internal consistency, unlike any of the previous models. The next generation, GPT-3, launched in 2020, was trained with 175 billion parameters. GPT-3 is a multi-purpose language tool that users can access without requiring them to learn a programming language or other computer tools. In November 2022, OpenAI released ChatGPT, which is a superior version of the company's earlier text generation models with the capability to generate humanlike prose.

After the success of ChatGPT (GPT 3.5), Open AI released GPT-4 in March 2023, which has multimodal capabilities. The model processes both image and text inputs for text generation. The model has a maximum token count of 32,768 capable of generating around 25,000 words as compared to GPT-3.5 which has 4,096 tokens context size. GPT-4 produces 40% more factual responses and its response rate for disallowed content is down by 82% as compared to previous models. (reported by OpenAI)

2. Google's Palm Models

Google AI, formerly known as Google Research, is the AI research and development arm of Google. It was unveiled at Google I/O 2018. Google has contributed many of the most significant papers in breakthroughs in modern machine learning.

Google's largest publicly disclosed model is its Pathways Language Model (PaLM) which has likely recently been rolled out in its Bard chatbot.

PaLM has been used as a foundation model in several Google projects including the instruction tuned PaLM-Flan, and the recent PaLM-E (the first "embodied" multimodal language model).

The pre-training of PaLM involved self-supervised learning drawing from a large text corpus that included multilingual web pages (27%), English books (13%), open-source code repositories, and source code from GitHub (5%), multilingual Wikipedia articles (4%), English news articles (1%), and other social media conversations (50%). PaLM excelled in 28 out of 29 NLP tasks in the few-shot performance, beating the prior larger models like GPT-3 and Chinchilla.

PaLM variants scale up to 540 billion parameters (vs GPT-3 at 175 billion) and trained on 780 billion tokens (vs GPT-3 300bn) — totalling around 8x more compute training than GPT-3 (but likely considerably less than GPT-4). PaLM was trained across multiple TPU v4 pods. Being a dense decoder-only Transformer model, PaLM is trained on two TPU V4 pods connected over a data center network and uses a combination of model and data parallelism. Researchers used 3072 TPU v4 chips in each pod, attached to 768 hosts. This large TPU configuration allows for efficient scale training without using pipeline parallelism. The Pathways system allows for scaling a model across Google's thousands of Tensor Processing Unit chips.

3. DeepMind's Chinchilla Model

DeepMind Technologies, founded in 2010, is a British AI research laboratory. It became a wholly owned subsidiary of Alphabet Inc., in 2015 after its acquisition by Google in 2014. DeepMind has created a neural network or a Neural Turing machine that tries to replicate the short-term memory of the human brain.

In 2016, DeepMind's AlphaGo program defeated a human professional Go player, and their program AlphaZaro defeated the most powerful programs in the games of Go and Shogi. The program acquired competence using reinforcement learning. In 2020, DeepMind's program AlphaFold started making advances in the problem of protein folding and by July 2022, it had predicted over 200 million protein structures. In April 2022, Flamingo, a single visual language model program capable of describing any picture, was launched. Three months later, in July 2022, DeepNash was announced; as a model-free multi-agent reinforcement learning system.

DeepMind developed a language model called Chinchilla AI in March 2022, which claimed to outperform GPT-3. A key breakthrough in the Chinchilla paper was that previous LLMs had been trained on too little data — for a given parameter size the optimum model should use far more training data than GPT-3. While more training data takes more time to gather, and leads to more training costs, achieving more capable models for a smaller parameter size has huge benefits for inference costs (the costs needed to run and use the finished model which scale with parameter size).

Chinchilla has 70B parameters (60% smaller than GPT-3) and was trained on 1,400 tokens (4.7x GPT-3). The average accuracy rate of Chinchilla AI is 67.5% on Measuring Massive Multitask Language Understanding (MMLU) and outperforms other large language model platforms like Gopher (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (300 parameters and 530B parameters) on a large range of downstream evaluation tasks.

4. Microsoft & Nvidia's Megatron Turing Model

Nvidia is a company that designs GPUs and APIs for data science and high-performance computing, and SoCs for mobile computing and the automotive market. The company is a leading supplier of AI hardware and software. Additionally, Nvidia's CUDA API enables the creation of massively parallel programs that leverage GPUs.

Developed by NVIDIA's Applied Deep Learning Research team in 2021, the Megatron-Turing model consists of 530 billion parameters and 270 billion training tokens. Nvidia has provided access via an Early Access program for its managed API service to its MT-NLG model.

Nvidia has made many of its LLM and Generative AI models and services available through its new DGX Cloud platform.

5. Meta's LlaMa Models

Meta AI, formerly known as Facebook Artificial Intelligence Research (FAIR), is an artificial intelligence laboratory that aims to share open-source frameworks, tools, libraries, and models for research exploration and large-scale production deployment. In 2018, they released the open-source PyText, a modeling framework focused on NLP systems. Then, in August 2022, they announced the release of BlenderBot 3, a chatbot designed to improve conversational skills and safety. In

November 2022, Meta developed a large language model called Galactica, which assists scientists with tasks such as summarizing academic papers and annotating molecules and proteins.

Released in February 2023, LLaMA (Large Language Model Meta AI) is a transformer-based foundational large language model by Meta that ventures into both the AI and academic spaces. The model aims to help researchers, scientists, and engineers advance their work in exploring AI applications. It will be released under a non-commercial license to prevent misuse, and access will be granted to academic researchers, individuals, and organizations affiliated with the government, civil society, academia, and industry research facilities on a selective case-by-case basis. The sharing of codes and weights allows other researchers to test new approaches in LLMs.

The LLaMA models have a range of 7 billion to 65 billion parameters. LLaMA-65B can be compared to DeepMind's Chinchilla and Google's PaLM. Publicly available unlabeled data was used to train these models, and training smaller foundational models require less computing power and resources. LLaMA 65B and 33B have been trained on 1.4 trillion tokens in 20 different languages, and according to the Facebook Artificial Intelligence Research (FAIR) team, the model's performance varies across languages. The data sources used for training included CCNet (67%), GitHub, Wikipedia, ArXiv, Stack Exchange, and books. LLaMA, like other large scale language models, has issues related to biased & toxic generation and hallucination.

6. Eleuther's GPT-Neo Models

Founded in July 2020 by Connor Leahy, Sid Black, and Leo Gao, EleutherAI is a non-profit AI research lab The organization has emerged as a leading player in large-scale natural language processing research, with a focus on interpretability and alignment of large models. Their mission is to ensure that the ability to study foundation models is not limited to a few companies, promoting open science norms in NLP, and creating awareness about capabilities, limitations, and risks around these models.

In December 2020, EleutherAI curated a dataset of diverse text for training LLMs called the Pile, which consisted of an 800GiB dataset. Subsequently, in March 2021, they released GPT-Neo models. EleutherAI also released GPT-J-6B in June 2021, which is a 6 billion parameter language model, making it the largest open-source GPT-3 like model at the time. Additionally, they combined CLIP with VQGAN to develop a free-to-use image generation model, which guided the foundation of Stability AI. EleutherAI also trains language models in other languages, such as Polyglot-Ko, which were trained in collaboration with the Korean NLP company TUNiB.

EleutherAI used Google's TPU Research Cloud Program, but by 2021, they took funding from CoreWeave. The company also uses TensorFlow Research Cloud for cheaper computing resources. In February 2022, EleutherAI released the GPT-NeoX-20b model, which became the largest open-source language model of any type at the time. In January 2023, the company was formally incorporated as a non-profit research institute.

EleutherAI's NLP model, GPT-NeoX-20B, is trained on 20 billion parameters using the company's GPT-NeoX framework and GPUs from CoreWeave. The GPT-NeoX-20B model has a 72% accuracy on LAMBADA sentence completion. When measured for zero-shot accuracy for Stem using Hendrycks Test Evaluation, it had an average of 28.98%. The model uses the Pile dataset for training and consists of data from 22 sources that falls under the following 5 categories: academic writing (Pubmed Abstracts and PubMed Central, arXiv, FreeLaw, USPTO Backgrounds, PhilPapers, NIH Exporter), web-scrapes and Internet resources (CommonCrawl, OpenWebText2, StackExchange, Wikipedia-English), prose (BookCorpus2, Bibliotik, Project Gutenberg), dialogue (Youtube subtitles, Ubuntu IRC, OpenSubtitles, Hacker News, EuroParl), and miscellaneous (GitHub, the DeepMind Mathematics dataset, Enron Emails).

GPT-NeoX-20B is publicly accessible and a pre-trained general-purpose autoregressive transformer decoder language model. It is a powerful few-shot reasoner with 44 layers and a hidden dimension size of 6144 and 64 heads. Additionally, it uses 1.1. Rotary Positional Embeddings instead of learned positional embeddings, as found in GPT models.

7. Cohere's XLarge

Model	Classify	Generate	Embed	Summarize
Default Model	$0.2 per 1000 Classifications	$15.0 per 1M Tokens	$0.4 per 1M Tokens	$15.0 per 1M Tokens
Custom Model	$0.2 per 1000 Classifications	$30.0 per 1M Tokens	$0.8 per 1M Tokens	-

Founded in 2019 by Aidan Gomez, Ivan Zhang, and Nick Frosst, Toronto-based Cohere specializes in natural language processing (NLP) models. Cohere has improved human-machine interactions and aided developers in performing tasks such as summarizing, classification, finding similarities in content, and building their own language models. Cohere's API helps users design tools for language comprehension and offers a backend toolkit for integration in multiple ways.

Cohere provides two types of large language models: Generation Language Models and Representation Language Models. The company uses a foundation model to train AI systems on large-scale data, enabling them to learn from new data to perform various tasks. Generative AI aims to develop human-like creations through coding, and Cohere competes with similar model providers like OpenAI and Anthropic, with the point of differentiation being the focus on serving enterprise users in incorporating generative AI. Cohere's goal is to make NLP accessible to all while building machines that are safe to use.

In September 2021, Cohere raised $40 million, and a few months later, in November 2021, Google Cloud announced its partnership with Cohere. The company intends to use Cloud's TPU for the development and deployment of its products, and Sagemaker by Amazon also gives access to Cohere's language AI. Cohere powers Hyperwrite, which helps in quickly generating articles. AWS has also announced a partnership with Cohere AI. To date, Cohere has raised $170 million, and with the ongoing rush of funding in AI platforms, the Canadian startup is expected to be valued at $6 billion.

Cohere is set to introduce a new dialogue model to aid enterprise users in generating text while engaging with the model to fine-tune the output. Cohere's Xlarge model resembles ChatGPT but provides developers and businesses with access to this technology. Cohere's base model has 52 billion parameters compared to OpenAI's GPT-3 DaVinci model, which has 175B parameters.

Cohere stresses on accuracy, speed, safety, cost, and ease of use for its users and has paid much attention to the product and its design, developing a cohesive model.

8. Anthropic AI's Claude

Model Family	Context Window	Prompt Pricing	Completion Pricing
Claude Instant	100,000 tokens	$1.63/ million Tokens	$5.51/ million Tokens
Claude-v1	100,000 tokens	$11.02/ million Tokens	$32.68/ million Tokens

Anthropic is an American AI startup and public benefit corporation founded in 2021 by Daniela Amodei and Dario Amodei, former members of OpenAI. The company specializes in developing AI systems and language models, with a particular focus on transformer architecture. Anthropic's research on the interpretability of machine learning systems covers fields ranging from natural language and interpretability to human feedback, scaling laws, reinforcement learning, and code generation, among others. The company stresses the application of responsible AI and presents itself as an AI safety and research company working towards building reliable, steerable, and interpretable AI systems.

By 2022, Google had invested nearly $400 million in Anthropic, resulting in a formal partnership between the two companies and giving Google a 10% stake in Anthropic. Outside backing amounted to $580 million, with total investments in Anthropic exceeding $1 billion to date.

Anthropic has developed a conversational large language model AI chatbot named Claude, which uses a messaging interface and a technique called constitutional AI to better align AI systems with human intentions. AnthropicLM v4-s3 is a 52-billion-parameter, autoregressive model, trained unsupervised on a large text corpus. The ten principles used by Anthropic are based on the concepts of beneficence, non-maleficence, and autonomy. Claude is capable of a variety of conversational and text-processing tasks, such as summarization, search, creative and collaborative writing, Q&A, and coding. It is easy to converse with, more steerable, and takes directions on personality, tone, and behavior.

Anthropic offers two versions of Claude — Claude (Claude-v1) and Claude Instant. Claude-v1 is a powerful, state-of-the-art high-performance model capable of handling complex dialogue, creative content generation, and detailed instructions. Claude Instant is lighter, less expensive, and much faster, making it suitable for handling casual dialogues, text analysis, and summarization. However, Claude is an expensive platform compared to ChatGPT.

Anthropic vouches for Claude to be an honest, helpful, and harmless AI system, and much less likely to produce harmful outputs than present chatbots, which have been known to be toxic, biased, use offensive language and hallucinate. According to Anthropic, Claude cannot access the internet and is designed to be self-contained and trained to avoid sexist, racist, and otherwise toxic outputs, along with preventing human engagement in illegal and unethical activities. However, compared to ChatGPT, Claude is poor at math and programming. Still, the platform has also been seen to hallucinate and provide dubious instructions. Another major concern is that it is possible to intrude upon Claude's built-in safety features through clever prompting.

The embargo on media coverage of Claude was lifted in January 2023, and a waiting list of users who wanted early access to Claude was released in February. Claude is now available and accessible to users through the Poe app by Quora. Also, Discord Juni Tutor Bot, an online tutoring solution, is powered by Anthropic. Additionally, Claude has found integration with Notion, DuckDuckGo, RobinAI, Assembly AI, and others.

9. AI21's Jurassic Models

AI21 Labs specializes in Natural Language Processing to develop generative AI models that can understand and generate text. The Tel Aviv-based startup was founded in 2017 by Yoav Shoham, Ori Goshen, and Amnon Shashua. AI21 has emerged as a rival to OpenAI. In 2019, the startup raised $9.5 million, and in October 2020; it launched Wordtune which was an AI-based writing app. AI21 Labs launched AI21 Studio and Jurassic-1 in August 2021. This was followed by Walden Catalyst investing $20 million in AI21 Labs in November, soon after which the company completed a $25 million series A round led by Pitango First. AI21 raised $64 million in the next round of funding. In January, AI21 Labs launched Wordtune Spices and Jurassic-2 in March 2023.

The Jurassic-1 model by AI21 Labs generates human-like texts and performs complex tasks like question answering, text classification, and others. The Jurassic-1 model comes in two sizes. Jurassic-1 Jumbo contains 178 billion parameters. The model uses a unique 250,000 token vocabulary and includes multi-word tokens, reducing the model's need to use a large number of tokens and thus improving the computational efficiency and reducing latency. Jurassic-1 allows developers to train custom versions of the model with just 50–100 training examples helping users to build customized applications and services. Jurassic-1 has been notably used by Latitude to scale production of its gaming world, by Harambee to create a custom chatbot to increase sign-ups for its youth employment programs, and by Verb to build a writing tool for authors.

The next iteration of Jurassic (Jurassic-2) is a highly customizable language model. It has comprehensive instruction tuning on proprietary data, which gives it advanced instruction following capabilities. The model supports languages like Spanish, French, German, Portuguese, Italian, and Dutch. Compared to the Jurassic-1 model, it has up to 30% faster response time, significantly reducing latency. Jurassic-2 has three sizes, with each one having a separate instruction-tuned version — Large, Grande, and Jumbo. Jurassic-2 helps users to build virtual assistants and chatbots and helps in text simplification, content moderation, creative writing, etc. Jurassic-2 also has zero-shot instruction capabilities. The model boasts of the most current knowledge and up-to-date database, with training being based on data updated in the middle of 2022, as compared to ChatGPT, which had closed its database by the end of 2021. Jurassic-2 comes with five APIs built for businesses that want specifically tailored generative AI features. The APIs include tools for paraphrasing, summarizing, checking grammar, segmenting long texts by topic, and recommending improvements. On Stanford's Holistic Evaluation of Language Models (HELM), Jurassic-2 Jumbo ranks second with an 86.8% win rate. Jurassic-2 is available for free till May 1st, 2023.

10. Baidu's ERNIE Model

Baidu, based in Beijing, is a prominent Chinese company that specializes in artificial intelligence. In 2019, Baidu launched a powerful AI language model named Ernie (Enhanced Representation through Knowledge Integration), which has been open-

sourced along with its code and pre-trained model based on PaddlePaddle.

Since its inception, Ernie has undergone significant improvements and can now execute a diverse array of tasks, such as language comprehension, language generation, and text-to-image generation. ERNIE was designed to enhance language representations by implementing knowledge masking strategies, such as entity-level masking and phrase-level masking. Baidu launched ERNIE 2.0 in July 2019, which introduced a continual pre-training framework. This framework incrementally builds and learns tasks through constant multi-task learning. ERNIE 3.0 was unveiled in early 2021 and introduced a unified pretraining framework that allows collaborative pretraining among multi-task paradigms. Unlike other models such as GPT-3, ERNIE 3.0 showcased task-agnostic zero-shot and few-shot learning capabilities and could be easily tailored for natural language understanding and generation tasks with zero-shot learning, few-shot learning, or fine-tuning. In late 2021, Baidu released ERNIE 3.0 Titan, a pre-training language model with 260 billion parameters that were trained on massive unstructured data.

Baidu developed ERNIE Bot, its latest large language model (LLM), and generative AI product. It is designed to serve as a foundational AI platform that can facilitate intelligent transformations in various industries, including finance, energy, media, and public affairs. Access to ERNIE Bot is currently limited to invited users, with the API expected to be available to enterprise clients through Baidu AI Cloud after application (as of March 16th).

Baidu aims to use the capabilities of ERNIE Bot to revolutionize its search engine, which holds the dominant position in China. Moreover, it is anticipated that ERNIE Bot will improve the operational efficiency of various mainstream industries, including cloud computing, smart cars, and home appliances.

Hardware and Cloud Platforms

Nvidia's H100 Tensor Core, their ninth-generation data center GPU, contains 80 billion transistors and is optimized for large-scale AI and high-performance computing (HPC) models. The A100, Nvidia's predecessor to the H100, is one of the best GPUs for deep learning. There is also Google's Tensor Processing Units (TPUs) which are custom-designed accelerator application-specific integrated circuits (ASIC) used for efficient machine learning workloads and are tightly integrated with TensorFlow, Google's machine learning framework.

Google Cloud Platform has opened availability of TPU v4 on Cloud, specifically designed to accelerate NLP workloads, and has also developed TPU v5 for use internally. Microsoft Azure also offers GPU instances powered by Nvidia GPUs, such as the A100 and P40, that can be used for various machine learning and deep learning workloads. Another key development is the partnership between Microsoft Azure and OpenAI, which has given OpenAI the resources to train both GPT-3 and GPT-4 that resulted in the availability of these models for developers in their applications through Azure's cloud infrastructure. AWS provides access to GPUs such as the Amazon Elastic Compute Cloud (EC2) P3 instances, which offer up to 8 Nvidia V100 GPUs with 5,120 CUDA cores and 300 GB of GPU memory. AWS has also developed its own chips for inference(Inferentia) and training (Trainium).

Several advanced models have been developed on these computing and cloud systems, including BERT, RoBERTa, Bloom, Megatron and the GPT family. BERT is one of the first pre-trained models that incorporated transformer architecture and resulted in state of the art scores in many NLP tasks. RoBERTa is a variant of BERT, trained on a much larger dataset with a more efficient training procedure. Lastly, Bloom is an open-access multilingual language model, containing 176 billion parameters and was trained on 384 A100–80GB GPUs.

The increasing availability of specialized hardware for NLP tasks represents a significant development in cloud computing programs. With the availability of these tools, companies can now train and run models that were previously impossible to build.

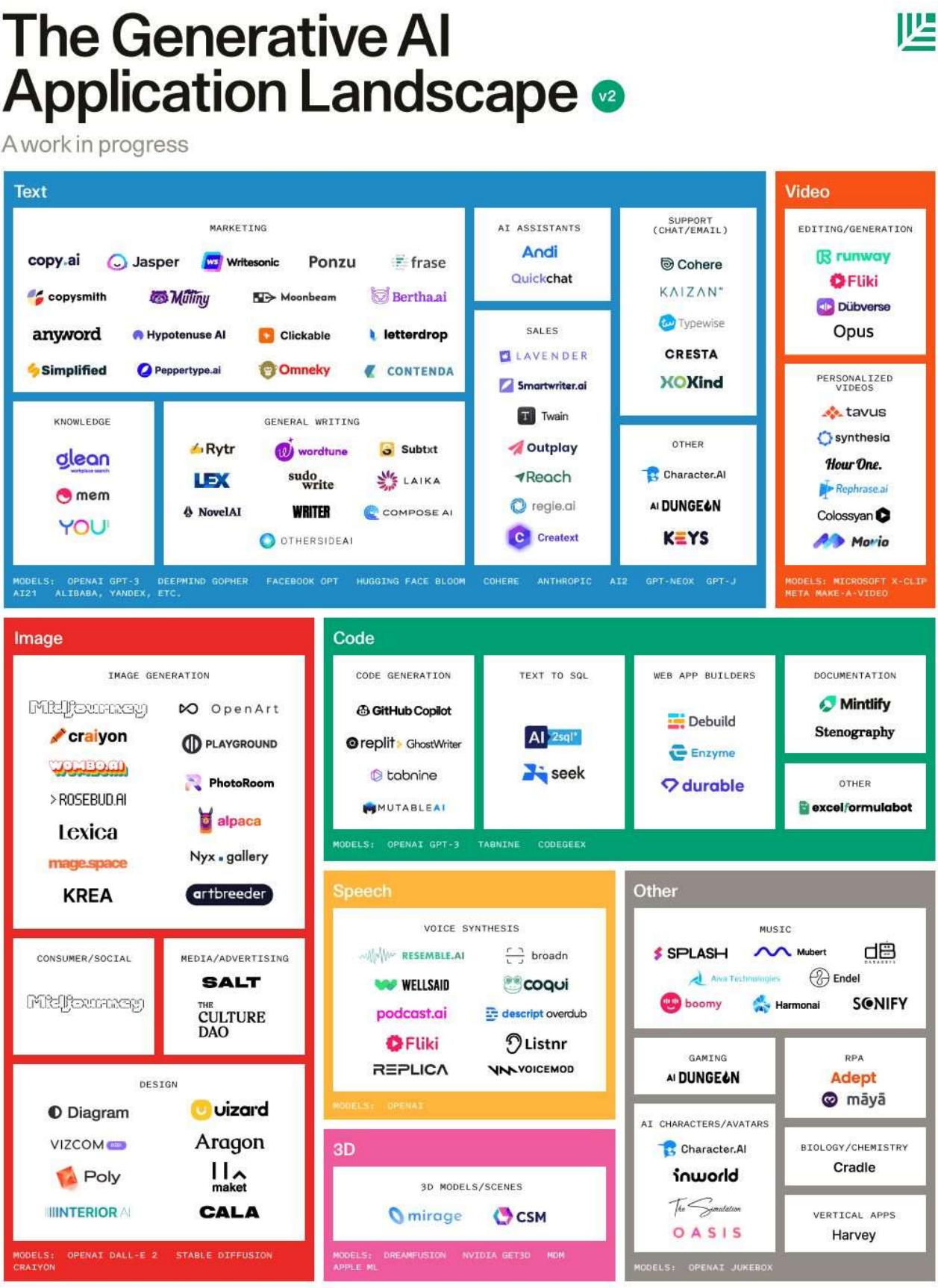
A note on Open Source

Open-source LLMs efforts have been progressing, both in terms of open data sets and open source models available for anyone to fine tune and use. The overall potential of open source models are very promising. They provide a more in-depth access to LLMs for everyone, not just by using an API. However there are definitely questions on the increased risks of models that haven't been aligned — and are more flexible to adapting for nefarious use cases such as misinformation.

AI efforts like Eleuther's "The Pile" and LAION's LAION-5B dataset have facilitated rapid progress in text and image modeling. Many companies and groups are also making foundational models accessible with open-source data sets, such as Big Science's Bloom model and the strategic partnership between Hugging Face and Amazon Web Services (AWS), which

increases the availability of open-source data sets and models hosted on Hugging Face. Stability AI also supports EleutherAI's work studying Large Language Models, while Laion's project involves crowdsourcing annotations for its OpenAssistant ChatGPT replication project. Additionally, Carper has developed open-source RLHF workflows ranging from human annotation with CHEESE to do RLHF training using trlX package.

Generative AI applied to other modalities



By some measures, consumer facing Generative AI has become the fastest growing technology trend of all time, with various models emerging for image, text, and code generation. For example, MidJourney's Discord has attracted around 13 million members for Image Generation, while ChatGPT has reportedly gained over 100 million users within a few months of release. Software development use cases have also seen a significant rise with over 1.2 million developers using GitHub Copilot's technical preview as of September.

1. Image Generation: Dall-E | MidJourney | Stable Diffusion | DreamStudio

Dall-E	MidJourney	Stable Diffusion	DreamStudio
$0.020 / image for 1024×1024 Resolution	$10/mo with 3.3 hr/month GPU Time	$9/mo & Training cost: $3/ Model	$10 for every 1000 generation credit

The combination of models, data, and computing has provided an incredible set of tools for working with images. OpenAI's DALL-E is an AI system that uses deep learning and transformer language models to generate digital images from natural language descriptions. It employs a decoder-only transformer model that models text and images as a single data stream containing up to 256 tokens for text and 1024 for images. The neural network then autoregressively models them. DALL-E is a 12-billion parameter version of GPT-3. The model uses a causal mask for text tokens and sparse attention for image tokens. DALL-E 2 is capable of producing higher-resolution images and uses zero-shot visual reasoning. It can create anthropomorphized versions, fill in the blanks, and transform existing images. However, DALL-E uses public datasets as training data, which can affect its results and often leads to algorithmic biases.

Midjourney is an artificial intelligence program developed by Midjourney, Inc., an independent research lab. The platform uses natural language descriptions to generate images, and users can create images by using Discord bot commands on the official Discord server. On March 16, 2023, beta version 5 was released. Users can generate images by typing the /imagine command followed by the prompt, and the bot generates four images, from which the user selects the image they want to upscale. Midjourney Inc. is also developing a web interface.

Stable Diffusion is an open source image model funded by Stability AI that generates images from text and performs tasks like inpainting, outpainting, and generating image-to-image translations. It uses a latent diffusion model supported by EleutherAI and LAION. It requires a minimum of 8GB VRAM making it independent of needing cloud services. Stable Diffusion 2.0 was released in November 2022 and trained on pairs of images and captions from LAION-5B and its subsets.

DreamStudio is the official online implementation and team interface API for Stable Diffusion, developed by Stability AI. DreamStudio and Stable Diffusion have slightly different interfaces even as they are applications of the same technology. The web app was launched in August 2022, replacing the free Discord bot. The web app offers better functionality and stability, using the Stable Diffusion algorithm to generate images based on the user's prompt. DreamStudio API Access has an access fee. One of the key features of DreamStudio is its support for negative prompting. It also allows users to overpaint, copy, modify, and distribute images for commercial purposes.

2. Audio Generation: Whisper | AudioGen | AudioLM

Whisper, developed by OpenAI, is a versatile automatic speech recognition system that supports multilingual speech recognition, speech translation, and language identification. It has been trained on 680,000 hours of multilingual and multitask supervised data using Python 3.9.9 and PyTorch 1.10.1, and the codebase is expected to be compatible with Python 3.8–3.10 and recent PyTorch versions. It deploys an encoder-decoder transformer model that uses 30-second chunks of input audio converted to log-Mel spectrograms, which are then passed to an encoder. The decoder predicts the corresponding text caption and intermixes special tokens to perform various tasks. Whisper provides an open-source model and inference codes for speech processing research and new application development. With nearly one-third of its dataset being non-English, Whisper outperforms the supervised state-of-the-art on CoVoST2 to English translation zero-shot.

Google's AudioLM is a pure audio model that uses language modeling to generate high-quality audio without annotated data. It generates speech continuations that preserve the identity, prosody, and accent of the speaker and recording conditions, and can also generate coherent piano music continuations. The model demonstrates long-term consistency in syntax, harmony, rhythm, and melody, and has the potential for extension to multilingual speech, polyphonic music, and audio events. AudioLM uses a hybrid tokenization scheme and a SoundStream neural codec to improve fidelity. The model

achieved a 51.2% success rate from human raters and an audio classifier with 98.6% accuracy was trained to detect synthetic speech generated by AudioLM. Currently, AudioLM is only available for research purposes and is not publicly available.

Meta's AudioGen AI converts text prompts into audio files. It is the audio parallel of image-generating AI like DALL-E. It uses a language AI model and approximately 4000 hours of training data to generate ambient sounds, sound events, and their composition. Additionally, it can extend existing audio to create rudimentary music. The quality of the audio output has been rated at 70% via Amazon's Mechanical Turk platform. However, AudioGen currently cannot sequence sounds through time, and the ownership rights of the generated audio are unclear.

3. Search Engines: Neeva | You

Neeva is an AI-powered search engine that provides ad-free and private searches. It achieves this through its in-house LLMs and search stack, while also blocking third-party website trackers and not sharing user information. Neeva's unique feature is its AI summaries, which provide synthesized answers backed by cited authority. It also allows users to search personal email accounts, calendars, and cloud storage platforms. This feature combines the best aspects of LLMs, like ChatGPT, with authority and timeliness. However, it only functions with question queries and has limitations on the free version (the premium plan is priced at $4.95/mo). Neeva has over 2 million users and local language versions in Germany, France, and Spain.

You.com is a California-based search engine that uses multimodal conversational AI to group web results into website categories sorted by user preferences. It was launched for public beta in November 2021 with a focus on privacy and personalization. It offers YouWrite, a text generator, and YouChat, a chatbot with community-built apps and blended LLMs. You.com does not collect users' personal information and offers personal and private search modes. The search results allow users to create content directly from the search results, building trust and reliability.

4. Code Generation: Copilot | Codex

GitHub Copilot is a tool that assists developers in programming by using AI to convert natural language into coding suggestions. It is powered by OpenAI Codex, which allows it to understand the developer's coding style and suggest context-specific solutions. When developers input their desired logic into the system, GitHub Copilot can generate code suggestions automatically. However, it is important to note that these suggestions are just that, suggestions, and it is up to the developer to decide whether to use them or not.

OpenAI Codex is a natural language processing model that is based on GPT-3 and can generate working code in multiple programming languages such as Python, JavaScript, and Ruby, among others. To train Codex, billions of lines of source code from public sources, as well as natural language data, including code from GitHub repositories, were used. It has a memory of 14KB for Python code and is a powerful, transformer-driven system that can effectively and efficiently fulfill developers' tasks.

5. Text Generation: Jasper

Jasper.AI is a subscription-based text generation model that requires minimal input from the user and searches the web to generate the desired output. It is particularly useful for generating short copy text where character limitations are important. The platform offers over 50 templates, including product descriptions, email subject lines, and Facebook headlines, among others. Additionally, it can help with generating ideas for blog posts and creating better outlines. However, Jasper.AI does have some drawbacks, such as the absence of fact-checking and citation of sources, which can lead to hallucinations. Additionally, learning the command input to achieve the desired output may take some time.

Conclusion

Generative AI is a revolutionary technology that has the ability to transform many aspects of our lives. Keep in mind that there are still challenges in developing these models such as massive datasets, compute power, high training cost, and accessibility. Studies have revealed that many large language models are not adequately trained. Additionally, smaller datasets are still crucial for enhancing LLM performance in domain-specific tasks. Compute cost optimization is also essential since generative models, especially large language models, are still expensive to both train and serve for inference. Big players in the industry are working on optimizing compute costs at every level.

Safety and security remain pressing concerns in the development of generative AI, and key players are incorporating human feedback to make the models safer from the outset. Open-source alternatives are also necessary to increase access to the next-generation LLM models for practitioners and independent scientists to push the boundaries forward.

Large Language Models | Artificial Intelligence | OpenAI | AI | Machine Learning