A Comprehensive Review of the Latest Advancements in Large Generative AI Models

Satyam Kumar(✉), Dayima Musharaf, Seerat Musharaf, and Anil Kumar Sagar

School of Engineering and Technology, Sharda University, Greater Noida, India
{2020540155.satyam,2020442645.dayima,
2020442638.seerat}@ug.sharda.ac.in, Anil.sagar@sharda.ac.in

Abstract. There has been an increase in big generative models like ChatGPT and Stable Diffusion over the last two years. These models are capable of a wide range of activities, including providing general answers and producing creative representations. They have a significant impact on a variety of businesses and society since they have the ability to transform established work roles. Generative AI may, for instance, convert text into images using the DALLE-2 model, 3D images using the Dreamfusion model, photos into text using the Flamingo model, and even text into video using the Phenaki model. While ChatGPT can translate text into other texts, the AudioLM model can translate text into audio. Text is converted into code by the Codex model and scientific texts using the Galactica model. Algorithms like AlphaTensor can also be developed through generative AI. This research seeks to present a thorough overview of the most important generative models that have recently been released and their impact on various industries. It also makes an effort to taxonomize these models in order to better comprehend their functions and applications.

Keywords: AI generative models · Text-to-image · Text-to-voice · Text-to-video · Text-to-code · Text-to-3D · Text-to-text · Text-to-Science · Image-to-text

1 Introduction

A specific type of artificial intelligence identified as "generative AI" is capable of producing new material as opposed to only analysing or processing data, as was the case with conventional expert systems. Knowledge bases and an inference engine that produced content using an if-else rule database made comprised expert systems.

But discriminator and generator now make up the two main parts of contemporary generative AI, which has greatly advanced in recent years. The discriminator can translate input data into a high-dimensional latent space after being trained on a corpus or dataset [1]. The generator, on the other hand, responds to a prompt by producing stochastic behaviour and unique material, even if the prompt is repeated. Depending on the methods employed, the learning process might be unsupervised, semi-supervised, or supervised.

It's critical to understand how generative AI models vary from systems that use predictive machine learning. Systems that use predictive machine learning only carry

R. N. Shaw et al. (Eds.): ICACIS 2023, CCIS 1920, pp. 90–103, 2023.
https://doi.org/10.1007/978-3-031-45121-8_9

out discrimination tasks to address classification or regression issues. On the other hand, generative AI models have the capacity to both recognise information and produce new information from the altered input data or prompt.

It's critical to understand how generative AI models vary from systems that use predictive machine learning. Systems that use predictive machine learning only carry out discrimination tasks to address classification or regression issues [2]. On the other hand, generative AI models have the capacity to both recognise information and produce new information from the altered input data or prompt.

Generative models' scale and ability to process enormous amounts of data are key features. It is now possible to feed generative models a vast amount of data, including the entirety of Wikipedia, Github, social networks, Google pictures, and more, because to breakthroughs in computing technology. Deep neural networks, transformers, and other models like variational autoencoders and generative adversarial networks have all emerged recently, making it possible to describe the complexity of this data without worrying about underfitting.

The high-dimensional probability distribution of words or images in a particular or generic domain can be modelled using these models. It is feasible to convert input data between different formats when combined with generative models that map the latent high-dimensional semantic space of language or images to a multimedia representation like text, audio, or video [3]. This opens up a wide range of application possibilities because a model may be taught to produce multimodal outputs in diverse formats, such as text, audio, or video, from different input formats.

In this paper, we intend to give a thorough review of the most widely used generative AI models and their effects on numerous sectors, including the arts and education. These sectors must change and adopt these methods because they can produce original artistic content and lengthy texts, continuing to add value. These models inspire artists and raise the standard of the content produced by instructors, not taking the place of human labour. The organisation of the paper is as follows: The primary generative models now in use are presented in a taxonomy first. We then examine the applications of each taxonomy category in depth. We wrap up by summarising the results and making recommendations for additional research. Note that we are not interested in the technical intricacies of these models' operation, but rather the applications of these models and the material they provide. We suggest using other references to gain a deeper understanding of deep learning and generative models.

2 Classification of Generative Artificial Intelligence Models

In this study, we have made an effort to taxonomize the present generative artificial models based on the primary mapping between input and output categories for multimedia. Figure 1 shows our results and emphasises 9 categories of models. Each of the models shown in Fig. 1 will have a detailed discussion in the section that follows. The emphasis of this manuscript is on the most recent developments in generative AI models, and all models mentioned have just recently been published, it is crucial to emphasise.

The deployment of these concepts has only involved six organisations. This is due to the fact that estimating the parameters of such models calls for a lot of processing power

and a highly skilled team of data scientists and engineers. Therefore, only businesses that worked in tandem with acquired startups and academic institutions were able to successfully implement generative AI models. After examining and presenting the most recent generative artificial intelligence models, we will go into further detail about each category in the taxonomy shown in Fig. 1 in the following section.

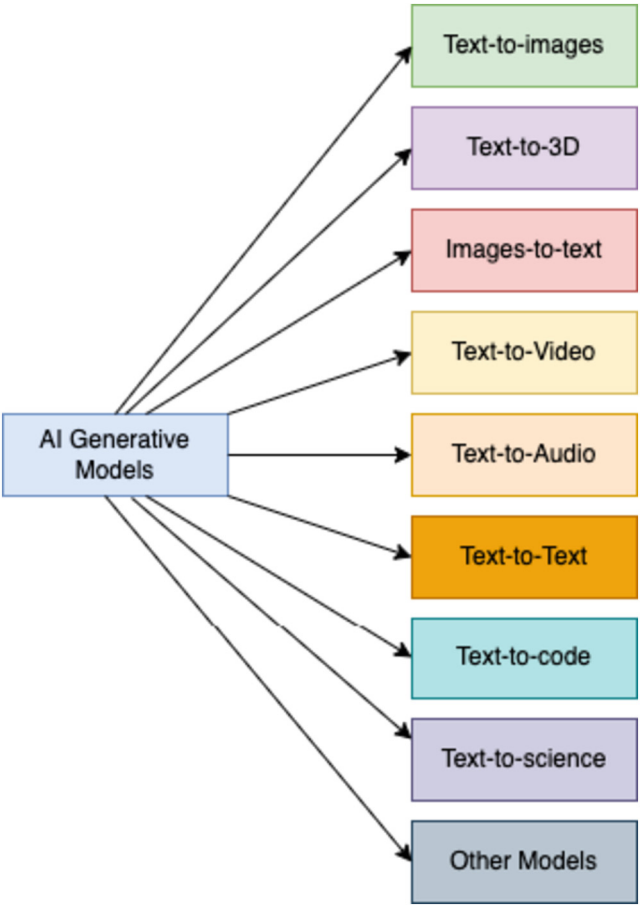

Fig. 1. A taxonomy depicting the classification of popular generative AI models based on their input and the format of the output they generate.

3 Categories of Generative AI Models

The nine categories depicted in Fig. 1 from the previous part will be further discussed in this section.

3.1 Text-to-Image Models

We'll look at models that accept a text prompt as input and produce an image to begin the analysis.

DALL·E 2. In response to a written prompt, DALLE 2, a generative model created by OpenAI, can create real-world visuals and artwork. Fortunately, this model can be accessed using the OpenAI API. As a result of utilising the CLIP neural network, DALLE 2 is able to combine ideas, attributes, and styles. Various (image, text) pairs were used to train the neural network known as CLIP (Contrastive Language-Image Pre-Training). With the help of instructions in natural language, CLIP can choose the most pertinent text excerpt given an image and has lately become a successful representation learner for images [4]. The desirable characteristics of CLIP embeddings include their amazing zero-shot capabilities, robustness to picture distribution shift, and fine-tuning to produce cutting-edge results. The CLIP image embedding decoder module is paired with an earlier model that creates potential CLIP image embeddings from a given text caption in order to obtain a whole generative model of images. Figure 2 displays an illustration of an image produced in response to a prompt.

Fig. 2. Image generated from the prompt "passing by street in cold winter while snowing at night in london street, digital art".

IMAGEN. It is a model that does text-to-image synthesis using huge transformer language models. Large language models that have been trained on text-only data are particularly good at encoding text for picture synthesis, which is one of the model's key conclusions. This model was made by Google, and their website has information on its API. The model has a good level of realism and image-text alignment, per Google's tests [5]. Google developed Drawbench, a collection of 200 prompts that facilitate the evaluation and comparison of text-to-image models, to assess the model. In order to translate text to a series of word embeddings, IMAGEN use a pretrained text encoder

like BERT. Then, these embeddings are mapped to images with progressively higher resolutions using a cascade of conditional diffusion models.

Stable Diffusion. Alternative generative AI models include Stable Diffusion, a latent-diffusion model created by the CompVis group at LMU Munich. The use of a latent diffusion model, which permits image alteration through operations in the model's latent space, distinguishes Stable Diffusion from other diffusion models. Stable Diffusion's website API can be used to access it. The model comprises of a text encoder and an age generator, the latter of which operates only in the latent space for faster diffusion than earlier models [6]. Figure 3 displays an illustration of an image produced using Stable Diffusion.

Fig. 3. Image generated from the prompt "frog jumping in rain, digital art".

Muse. The Muse model is a text-to-image converter that produces stunning images while being more effective than diffusion or autoregressive models. This is because it requires fewer sample iterations and uses discrete tokens [7]. Muse is more effective than an autoregressive model due to parallel decoding; it is 10 times faster at inference time than Imagen-3B or Parti-3B and 3 times faster than Stable Diffusion v1.4, despite the fact that both models operate in a VQGAN's latent space.

3.2 Text-to-3D Models

Models that turn text cues into 2D graphics were covered in the section before. However, some sectors, like the gaming industry, demand the creation of 3D images. We will provide a quick description of two text-to-3D models in this section: DreamFusion and Magic3D.

DreamFusion. It is a Google Research model that seeks to produce 3D graphics from text cues. To do this, the model distils a loss from a 2D diffusion model and substitutes the prior CLIP approaches with a pretrained 2D text-to-image diffusion model. This implies

that samples are produced using the diffusion model as a loss in a general continuous optimisation problem. Creating 3D models that resemble decent photos when rendered from random angles is difficult because sampling in parameter space is substantially more difficult than sampling in pixel space [8]. It tackles this issue by using a differentiable generator rather than only focusing on sampling pixels. This enables the model to produce 3D models that from various perspectives resemble appealing photographs. It's important to note that the DreamFusion website allows you to view the entire animated image.

Magic3D. The Dreamfusion text-to-3D model's drawbacks of a protracted processing time and poor image quality have been addressed by NVIDIA Corporation's Magic3D text-to-3D model [20]. A sparse 3D hash grid structure is used to accelerate a low-resolution diffusion prior that Magic3D builds using a two-stage optimisation framework. A textured 3D mesh model is produced as a result of this procedure, which is further optimised using a quick differentiable render [9]. Magic3D outperforms DreamFusion in human evaluations, with 61.7% of respondents saying they prefer it. The quality of the generated 3D shapes is far superior than that of DreamFusion's, both in terms of geometry and texture.

3.3 Image-to-Text Models

The task of producing a textual description of a picture, which is the inverse mapping of the text-to-image synthesis models previously addressed, is the main topic of this section. This section examines Flamingo and VisualGPT, two models that carry out this duty in addition to others.

Flamingo. It is a Visual Language Model made by Deepmind that only requires a small amount of learning to carry out a variety of vision and language tasks when given a limited number of input/output samples. It is made up of visually conditioned autoregressive text generation models that accept as input a string of text tokens interspersed with images and/or videos and generate text as the result [10]. Flamingo produces a text response in response to a query that includes a photo or video. The model combines a huge language model that executes a fundamental form of reasoning with a vision model that analyses visual scenes. On a sizable corpus of textual data, the language model is trained.

VisualGPT. It is a model for image captioning created by OpenAI that makes use of the GPT-2 language model's expertise. To close the semantic gap between various modalities, the model includes a unique encoder-decoder attention mechanism with an unsaturated rectified gating function [11]. The fact that VisualGPT uses less data than other image-to-text algorithms is one of its main advantages. This may enable rapid data curation, descriptions of unusual objects, and applications in specialised fields. Additionally, this model's API is accessible on GitHub.

3.4 Text-to-Video Models

The capacity to create images from text has been proven in the previous subsections. Given this, it makes sense to produce videos, which are essentially a series of images,

from text. We will talk about two models that are capable of doing this in this section: Phenaki and Soundify.

Phenaki. Phenaki is a video synthesis model created by Google Research that can produce lifelike films in response to textual cues. This model's API is more broadly accessible because it is accessible from GitHub. The ability of Phenaki to generate films from open domain time-variable cues makes it special. It was trained on both a sizable dataset of image-text pairs and a smaller dataset of video-text examples in order to overcome difficulties with data scarcity. Because text-video datasets are smaller and have fewer inputs than image-text datasets, this approach was able to generalise beyond the capabilities of video datasets. Due to computing capacity for videos of varied length, there are restrictions, nevertheless. The C-ViViT encoder, training transform, and video generator are the three parts of the model. Videos are compressed by the encoder, and the initial tokens are converted into embeddings before being passed via a spatial and temporal transformer [12]. The output of the spatial transformer is then activate-free, single-linear projected back to pixel space. This method creates different, temporally consistent films based on free domain prompts, even when the prompt is a novel concept assemblage. The model can produce videos up to several minutes in length, while being trained on 1.4 s videos.

Soundify. Runway created a method called Soundify to help professionals with video editing discover and match the right sounds. The system makes use of high-quality sound effect libraries and the zero-shot image classification capabilities of CLIP, a neural network. Specifically, classification, synchronisation, and mix make up its three main components. Sound effects are matched to videos in the classification phase by categorising the sound emitters present. [13] The video is divided based on absolute colour histogram distances, which helps to reduce the number of unique sound emitters. In the synchronisation phase, intervals are found by comparing the labels of the effects with each frame and spotting repeated matches above a predetermined threshold. Effects are divided into one-second segments in the mix section, which are subsequently combined using crossfades.

3.5 Text-to-Audio Models

We discovered in the previous subsection that non-structured data formats other than photos are also significant. In many instances, like movies and music, audio is also essential. So, in this subsection, we'll look at three models that accept text as input and output audio.

AudioLM. By converting the input audio into a series of discrete tokens and treating the synthesis of audio as a language modelling task in this representation space, Google has created the AudioLM model, which produces high-quality audio with long-term consistency. The model is able to provide realistic and coherent continuations when given brief instructions because it was trained on vast corpora of unprocessed audio waveforms. Despite not having been trained with any symbolic representation of music, the model can even be extended beyond speech to produce logical piano music continuations. This

model's API is available on GitHub. It might be difficult to provide great audio quality while demonstrating consistency since audio signals contain several abstraction scales.

To overcome this difficulty, AudioLM combines recent developments in neural audio compression, self-supervised representation learning, and language modelling. Raters listened to a 10-s clip and identified whether it was human speech or a synthetic continuation to assess the performance of the model. Based on 1000 ratings, 51.2% of the ratings were correctly classified, which is not statistically significant when labels are chosen at random [14]. This shows that it is impossible for humans to tell the difference between artificial and actual audio samples produced by the model.

Jukebox. This article discusses an OpenAI model for making music that can produce songs with singing in the raw audio domain. On GitHub, you can access the model's API. The goal of this technique is to produce music directly as raw audio, in contrast to earlier text-to-music models that produced symbolic piano-roll representations. However, due to the high dimensionality of raw audio and its high long-range dependencies, learning the high-level semantics of music is challenging. The model uses a hierarchical VQ-VAE architecture to compress the audio into a discrete space while maintaining the most information possible with a unique loss function in order to address this. The model's VQ-VAE contains 5 billion parameters and was trained over three days on 9-s audio clips using a dataset of 1.2 million English songs from LyricWiki [15]. The model is able to produce music in a variety of genres, including jazz, rock, and hip-hop.

Whisper. An audio-to-text converter model that was created by OpenAI is capable of carrying out a number of tasks, including language identification, translation, and multilingual speech recognition. Its API is accessible on the GitHub website, similar to that of other models. The primary difficulty for a voice recognition system is to function well in a variety of settings without necessitating supervised decoder fine-tuning for each deployment distribution. However, the absence of a top-notch pre-trained decoder makes this difficult to accomplish. The model is trained using a sizable dataset of 680,000 h of labelled audio data, which is acquired from the internet to cover a wide distribution of audio from various contexts, recording setups, speakers, and languages in order to solve this. To avoid degrading the model, the model makes sure that the dataset only contains human voice and leaves out machine learning voice. The architecture used by the model, an encoder-decoder transformer, has been shown to scale consistently.

3.6 Text-to-Text Models

The models mentioned so far concentrate on transforming unstructured data into different formats. To execute duties like responding to common questions, it may occasionally be necessary to transform text into another text. The four models listed below are examples of this kind; they handle text inputs and produce text outputs in order to satisfy specific needs.

ChatGPT. OpenAI created ChatGPT, a sophisticated conversational AI model. It is a powerful model that can interact in a conversational manner with users, providing natural language answers to questions and to follow-up inquiries. Similar to earlier language

models like GPT-2 and GPT-3, the model is based on a transformer architecture, allowing it to process enormous volumes of text and produce high-quality responses.

The capacity of ChatGPT to use reinforcement learning to learn from user comments is one of its distinctive features. Human instructors provided conversational scenarios where they took on the roles of both the user and the AI assistant during the training process. In order to assist the model, learn and develop, the human trainers would subsequently provide feedback on the model's responses to those conversations.

The training procedure for ChatGPT combines reinforcement learning with supervised fine-tuning. In the beginning, the model is trained using supervised fine-tuning, in which human trainers provide conversational examples, and the model learns to generate responses based on those examples [14]. The model is then improved further through reinforcement learning, where it receives comments from human trainers on its generated responses and learns to become more sociable.

ChatGPT is able to solve simple mathematical equations and generate code in addition to being able to carry on a conversation. This makes it a flexible tool for a variety of applications, such as chatbots, personal assistants, and customer support. Overall, ChatGPT is a promising tool for enhancing future human-AI interactions and represents a substantial leap in conversational AI (Figs. 4 and 5).

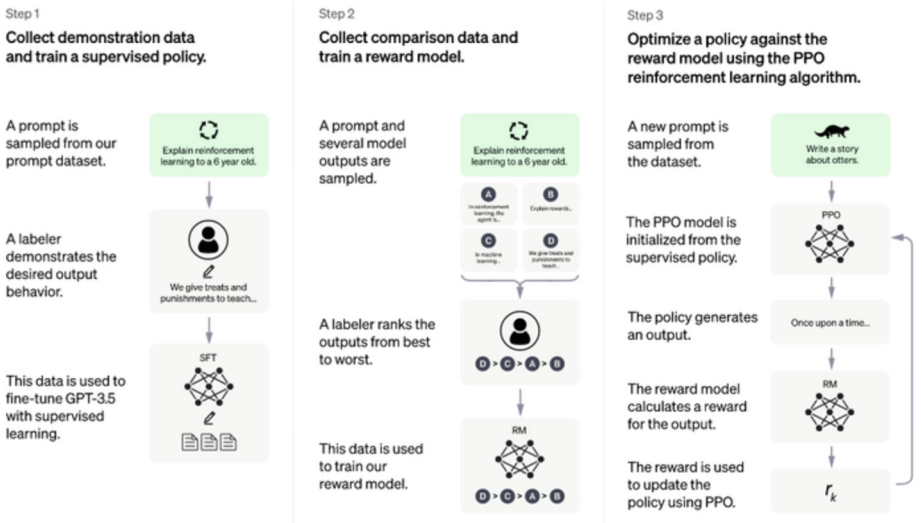

Fig. 4. Training steps of ChatGPT, combining supervised learning with reinforcement learning.

LaMDA. LaMDA is a specialised neural language model based on transformers that is intended for dialogue applications. It was particularly trained on dialogues, unlike other language models, and it has up to 137B parameters. One of the largest pre-trained language models, the model was pre-trained using 1.56T words of public dialogue data and web content. The model can be fine-tuned to increase its level of safety and factual

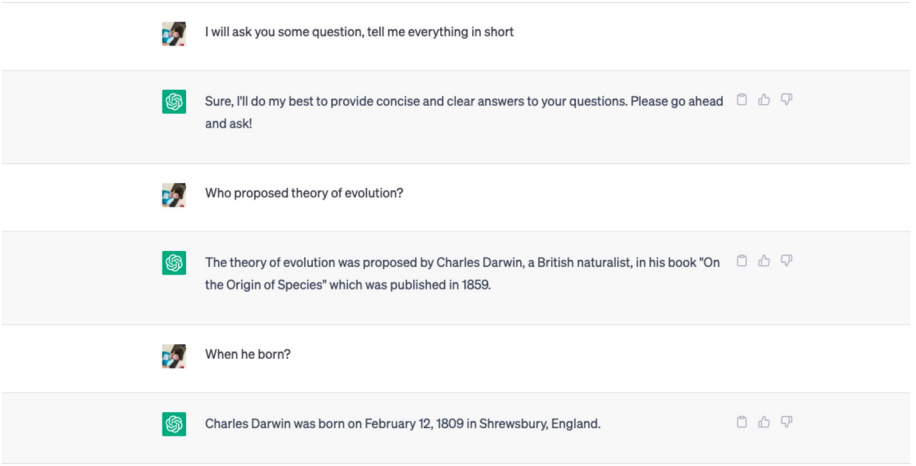

Fig. 5. Example of a dialog made with ChatGPT.

correctness. LaMDA is well appropriate for model scaling since it makes use of the Transformers' capacity to handle long-term dependencies in text.

The model uses a single architecture to carry out a number of tasks: it creates a number of potential responses to a prompt, eliminates any that are hazardous or improper, and then bases the remaining responses on an outside knowledge source to assure their accuracy [16]. The highest-quality response for the particular question is then chosen by re-ranking the grounded responses. The model is a useful tool for a range of applications since it can manage complex dialogues with various responses.

PEER. The "Collaborative Writing Assistant," a collaborative language model created by Meta AI research, is intended to encompass the complete writing process. The four steps of the model are Plan, Edit, Explain, and Repeat. Until the text is in a satisfactory state and doesn't need any more updates, these steps are repeated.

The concept enables the division of the writing of a paper task into numerous simpler subtasks. The model also enables human interaction at any time and permits model steering in any direction.

The strategy is self-training, utilising models to fill in missing data and then training other models on this synthetic data. The model is mostly trained using Wikipedia edit histories. A retrieval method that occasionally fails to make up for the fact that comments are frequently loud and devoid of citations is one drawback.

To create a sequence of texts, the entire process of developing a plan, gathering information, editing it, and presenting it can be repeated several times. A DeepSpeed transformer is employed during the training of this model. The Collaborative Writing Assistant is a promising tool for group writing that could help authors at every stage of the writing process.

Meta AI Speech from Brain. To assist those who are unable to communicate normally through speech, typing, or gestures, Meta AI has created a model. Such individuals have in the past been forced to rely on invasive brain-recording methods that demand

neurosurgical operations. This new model seeks to directly translate language from noninvasive brain recordings, offering a safer, more scalable method that may help a much larger population. The noise, individual variances in each person's brain, and the placement of the sensors provide a significant obstacle to the proposed strategy, though.

In order to correlate noninvasive brain recordings and speech sounds as precisely as feasible, the model is trained via contrastive learning. To be more precise, volunteers' listening to audiobooks is utilized to identify the complex representations of speech in their brains using a self-supervised learning model called wave2vec 2.0. Electroencephalography (EEG) and magnetoencephalography (MEG) are the two noninvasive techniques used to monitor neural activity. Four open-source datasets representing 150 h of recordings of 169 volunteers listening to audiobooks are the source of the training data.

The brain model is a conventional deep convolutional network with residual connections that incorporates EEG and MEG measurements. The brain activity of the individuals is represented by these recordings. The model includes a brain model for MEG data as well as a speech model for sound.

The study's findings suggest that a number of algorithmic elements were helpful to decoding performance, and an examination reveals that the algorithm gets better as the number of EEG and MEG recordings rises. The study shows that despite noise and data unpredictability, self-supervised trained AI can decipher perceived speech. Though the study primarily focuses on speech perception, expanding this research to include speech production is the ultimate objective. Despite this drawback, the research has positive implications for people who struggle to communicate in conventional ways.

3.7 Text-to-Code Models

While there are many models that can be used with text written in natural language, it's important to remember that not all text has the same syntax, particularly when it comes to programming code. Programming requires the translation of concepts into code, and tools like Codex and Alphacode can be helpful in accomplishing this process.

Codex. The AI system Codex, created by OpenAI, can translate language into code, making it a useful tool for programming chores. This all-purpose programming approach is intended to assist programmers in decomposing large issues into smaller, easier-to-manage ones, which can then be mapped to pre-existing code libraries, APIs, or functions. Coding's second phase is frequently the most time-consuming, and this is where Codex excels. It was trained on 179 GB of distinct Python files under 1MB that were gathered from GitHub's open-source software repositories in May 2020. GPT-3, which has strong natural language processing capabilities, provides the basis for Codex's fine-tuning [18]. On the website for OpenAI, one can access Codex's demo or API to use it.

Alphacode. Although some language models can produce code, they perform poorly when tested against challenging, uncharted problems. The language model Alphacode was created for code creation and can handle more complex reasoning. A combination of elements, including a sizable and effective transformer-based architecture, a large-scale model sampling strategy, and a huge dataset for training and evaluation, is the secret to its success.

Code from GitHub repositories totaling 715.1 GB makes up the dataset used by Alphacode for training, which is significantly greater than the dataset used by Codex for pre-training. A dataset from the Codeforces platform is also used to fine-tune the model. The coding competition platform Codeforces offers a useful data set for model validation and performance enhancement.

In contrast to decoder-only models frequently employed in other code generation systems, the architecture of Alphacode employs an encoder-decoder transformer paradigm, allowing for bidirectional description and enhanced flexibility [19]. The model makes use of both shallow and deep encoders to increase efficiency. Additionally, multi-query attention is used to lower sampling costs.

3.8 Text-to-Science Models

With the development of the Galactica and Minerva models, the application of generative AI is expanding to include scientific literature as well. Investigation of the earliest initiatives towards automating scientific text generation is crucial, despite the fact that there has been little development in this area.

Galactica. Galactica is a new huge AI model created by Papers with Code and Meta AI with the goal of autonomously organising scientific content. The model's capacity to train for numerous epochs without overfitting, enhancing upstream and downstream performance with repeated tokens, is one of its key advantages. All data is processed in a standard markdown format to combine knowledge from various sources, ensuring an effective strategy. The model uses a specific token to predict citations, and its accuracy rises with scale and improved citation distribution [20]. Additionally, the model is capable of multi-modal tasks like protein sequences and SMILES chemical formulas. Across all model sizes, Galactica uses a transformer architecture with GeLU activation that is only used in a decoder setup.

Minerva. A language model called Minerva was created primarily to answer mathematical and scientific problems using a sequential line of reasoning. The model may generate models at scale and use the finest inference approaches because the training data is carefully gathered to concentrate on quantitative reasoning issues. Without using external tools like calculators, Minerva generates solutions step-by-step using math and symbolic manipulation. This is a novel method of problem-solving.

3.9 Other Models

There are more models that are worthwhile taking into consideration in addition to the generative AI models already described. Deepmind's Alphatensor, which was created, is capable of finding new algorithms. For instance, it can produce more effective matrix multiplication algorithms, which is crucial for enhancing calculations in neural networks and scientific computing routines. The method makes advantage of deep reinforcement learning, and the AlphaTensor agent is trained to play a single-player game that entails locating tensor decompositions in a finite factor space. Synthetic training games are used by AlphaTensor to leverage symmetries using a specialised neural network architecture.

GATO, a single generalist agent that functions as a multi-modal, multi-task, and multi-embodiment generalist policy, is another Deepmind model. The same network may carry out many tasks including playing Atari, captioning photos, talking, stacking blocks, and more with the same weights. The requirement for manually creating policy models with their unique inductive biases is reduced when using a single neural sequence model for all tasks, while the quantity and variety of training data are also increased. This universal agent is effective at a variety of activities, and it can be modified with little additional data to be effective at an even greater variety of tasks.

As demonstrated by ChatBCG, other published generative AI models can similarly produce human motion or slides using ChatGPT as a stand-in model.

4 Conclusions and Further Work

The amazing capabilities of generative AI in tasks like text-to-image and text-to-audio, which demonstrate its inventiveness and personalisation, are highlighted in this research. Additionally, text-to-science and text-to-code tasks' accuracy shows the potential for generative AI to optimize both creative and non-creative tasks, which could have a significant positive impact on economies.

Although generative AI models have impressive capabilities, there are still a number of restrictions and difficulties that need to be resolved. The lack of datasets is one of the biggest problems, especially for difficult jobs like text-to-science or text-to-audio. Finding appropriate data takes time, and some models' massive dataset requirements make training them much more challenging. Another concern is that these models are limited in their capacity for innovation and adaptability since they struggle to handle issues that fall outside the purview of their training data.

The extensive computing required to run these models presents another difficulty. It can take several days or even weeks to train them because it demands sophisticated computational resources. Additionally, there is a chance that the training data will be biased, which could impact the models' precision and dependability. Although some models, like Galactica, make an effort to address this problem through bias reduction techniques, it is still a major obstacle for generative artificial intelligence.

The Minerva model, among other recent innovations, has demonstrated encouraging results in terms of comprehending and piecemealing equations. This represents a significant advance because one of the key drawbacks of these models is that they do not fully comprehend the tasks at hand. However, some models, like text-to-video, still struggle with accuracy because it's difficult to create accurate and realistic videos.

These models are also fraught with ethical issues, particularly in light of text-to-video technology's potential for producing deep fakes. As a result, it's important to limit the usage of these models and make sure their ethical implications are carefully taken into account.

Last but not least, we have only just begun to fully understand the potential and function of generative artificial intelligence. Given that each model has unique advantages and disadvantages, comparisons between them, such as that between Google and ChatGPT3, are not entirely accurate. It is crucial to be conscious of these restrictions and seek to enhance the models in the upcoming years.

References

1. Bhavya, B., Xiong, J., Zhai, C.: Analogy generation by prompting large language models: a case study of instructgpt. arXiv preprint arXiv:2210.04186 (2022)
2. Budzianowski, P., Vulic, I.: Hello, it's gpt-2–how can i help you? Towards the use of pretrained language models for task-oriented dialogue systems. arXiv preprint arXiv:1907.05774 (2019)
3. Chang, H., et al.: Muse: text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
4. Borsos, Z., et al.: AudioLM: a language modeling approach to audio generation. arXiv preprint arXiv:2209.03143 (2022)
5. Balaji, Y., et al.: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
7. Kim, J.-H., Kim, Y., Lee, J., Yoo, K.M., Lee, S.-W.: Mutual information divergence: a unified metric for multimodal generative models. arXiv preprint arXiv:2205.13445 (2022)
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
9. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
10. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
11. Chowdhery, A., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
12. Zhou, Q., et al.: A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv preprint arXiv:2302.09419 (2023)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
14. Lin, S., Hilton, J., Evans, O.: Truthfulqa: measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)
15. Rajawat, A.S., Bedi, P., Goyal, S.B., Shaw, R.N., Ghosh, A.: Reliability analysis in cyber-physical system using deep learning for smart cities industrial IoT network node. In: Piuri, V., Shaw, R.N., Ghosh, A., Islam, R. (eds.) AI and IoT for Smart City Applications. SCI, vol. 1002, pp. 157–169. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7498-3_10
16. Thoppilan, R., et al.: Lamda: language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
17. Pant, P., et al.: Study of AI and ML Based Technologies used in international space station. Glob. J. Innov. Emerg. Technol. **1**(2) (2022). https://doi.org/10.58260/j.iet.2202.0102
18. Carlini, Liu, Y., Daume III, H., Erlingsson, U., Kohno, T., Song, D.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21) (2021)
19. Madaan, A., Zhou, S., Alon, U., Yang, Y., Neubig, G.: Language models of code are few-shot commonsense learners. arXiv preprint arXiv:2210.07128 (2022)
20. Taylor, R., et al.: Galactica: a large language model for science. arXiv preprint arXiv:2211.09085 (2022)