ChatGPT is not all you need. A State of the Art Review of large Generative AI models

Roberto Gozalo-Brizuela, Eduardo C. Garrido-Merchán

Quantitative Methods Department, Universidad Pontificia Comillas, Madrid, Spain
`201905616@alu.comillas.edu, ecgarrido@icade.comillas.edu`

Abstract. During the last two years there has been a plethora of large generative models such as ChatGPT or Stable Diffusion that have been published. Concretely, these models are able to perform tasks such as being a general question and answering system or automatically creating artistic images that are revolutionizing several sectors. Consequently, the implications that these generative models have in the industry and society are enormous, as several job positions may be transformed. For example, Generative AI is capable of transforming effectively and creatively texts to images, like the DALLE-2 model; text to 3D images, like the Dreamfusion model; images to text, like the Flamingo model; texts to video, like the Phenaki model; texts to audio, like the AudioLM model; texts to other texts, like ChatGPT; texts to code, like the Codex model; texts to scientific texts, like the Galactica model or even create algorithms like AlphaTensor. This work consists on an attempt to describe in a concise way the main models are sectors that are affected by generative AI and to provide a taxonomy of the main generative models published recently.

1   Introduction

Generative AI refers to artificial intelligence that can generate novel content, rather than simply analyzing or acting on existing data like expert systems [29]. In particular, expert systems contained knowledge bases and an inference engine that generated content via an if-else rule database. However, modern generative artificial intelligence contain a discriminator or transformer model trained on a corpus or dataset that is able to map the input information into a latent high-dimensional space and a generator model, that is able to generate an stochastic behaviour creating novel content in every new trial even from the same prompts as an input, performing unsupervised, semi-supervised or supervised learning, depending on the particular methodology. Regarding the created content by the model, generative artificial intelligence models are different from predictive machine learning systems, that merely perform a discrimination behaviour, solving classification or regression problems. In particular, these models are able to discriminate information and generate information of the transformed input information, or prompt.

The key aspect about generative models is that their architecture and the data that they have been fed is enormous. For example, it is possible now to

estimate the parameters of the model by feeding it the contents of the whole Wikipedia, Github, social networks, Google images and more. Despite being fed with an enormous size of data, thanks to the rise of computing we can design deep neural networks [23], transformers [27] and other models such as generative adversarial networks [10] or variational autoencoders [29] whose capacity is able to model the complexity of the data, without suffering from underfitting. As they are able to modelize the high-dimensional probability distribution of language or photos of a concrete or general domain, if they are complemented by generative models that map the latent high-dimensional semantic space of language of photos to a multimedia representation of text, audio or video we can map any input format like texts to any output format like video. In this sense, applications of this technology are endless, in the sense that we can train a model to generate genuine different multimedia formats as video, audio or text from different multimedia input formats, as for example, text.

We believe that it is necessary to provide a state-of-the-art review on the most popular generative AI models as they are revolutionizing several industries like the art industry [3] or universities [21, 36]. As models are now able to generate genuine artistic content or large texts answering a prompt, these two industries and other ones that we will detail throughout this manuscript will need to readapt their activity to continue providing value. In this sense, generative AI models will not replace humans but enhance our content, being an inspiration for artists or improving the content generated by professors. In order to provide information for a professional working in any industry that can be benefited by these models we have made the organization of the paper as the following one. First, we will provide a taxonomy of the main generative models that have appeared in the industry. Then, the following sections will analyze each of the categories of the taxonomy. We then include a section about ethical issues of the mentioned generative models. Finally, we finish the manuscript with a conclusions and further work section. We do not study the technical aspects of every model, such as transformers in detail as our purpose in this review is on the applications of the models and the content that they generate but not on how they work. For a detailed explanation of deep learning models and generative models we recommend other references [23, 29].

2   A Taxonomy of Generative AI models

Before analyzing each model in detail, we have tried to organize the current generative artificial models into a taxonomy whose categories represent the main mappings between each multimedia input and output type of data. The result is the one that we have illustrated in Figure 1. We have discovered a total of 9 categories, where each of the models that appear in Figure 1 will be described in detail in the following section. Each of the covered models has been published recently, as we illustrate in Figure 2, as our main concern in this manuscript is to describe the latest advances in generative AI models.
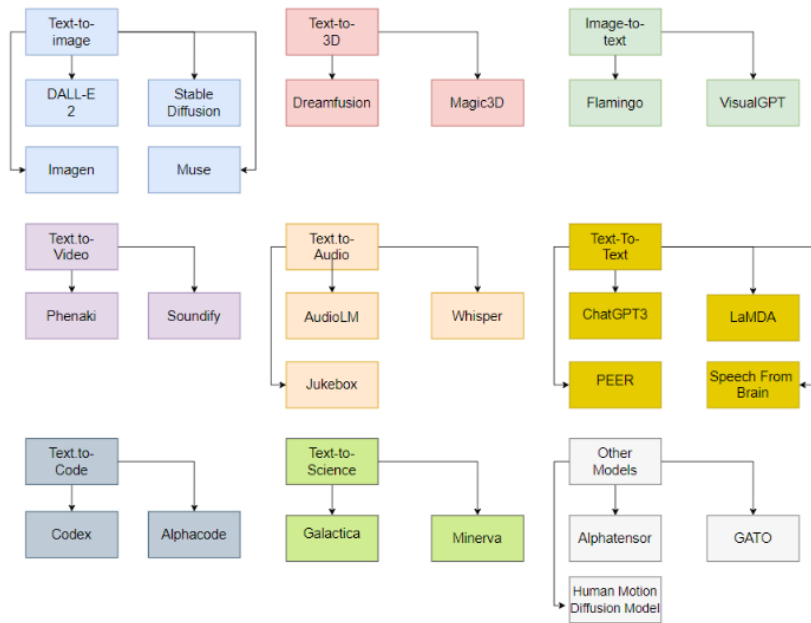
Fig. 1. A taxonomy of the most popular generative AI models that have recently appeared classified according to their input and generated formats.
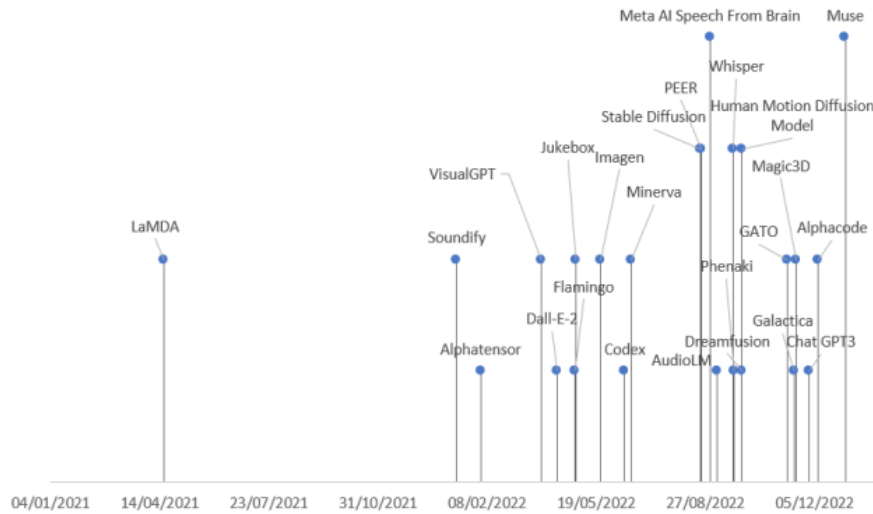

Fig. 2. Covered models by date of release. All models were released during 2022 except LaMDA, which was released in 2021 and Muse, in 2023.

Interestingly, only six organizations are behind the deployment of these models, as we illustrate in Figure 3. The main reason behind this fact is that in order to be able to estimate the parameters of these models, it is mandatory to have an enormous computation power and a highly skilled and experienced team in data science and data engineering. Consequently, only the companies shown on Figure 3, with the help of acquired startups and collaborations with academia, have been able to successfully deploy generative artificial intelligence models.

Now that we have provided and analyzed the latest generative artificial intelligence models, the following section will cover each of the categories of the taxonomy presented in Figure 1 in detail.

3   Generative AI models categories

In this section we will cover in detail the nine categories described in Figure 1 of the previous section. For every category, we illustrate the details of the models shown in Figure 1.

3.1   Text-to-image models

We begin the review by considering the models whose input is a text prompt and whose output is an image.

DALL·E 2 : DALL·E 2, created by OpenAI, is able to generate original, genuine and realistic images and art from a prompt consisting on a text description [11]. Luckily, it is possible to use the OPENAI API to get access to this model. In particular, DALL·E 2 manages to combine concepts, attributes and different styles. In order to do so, it uses the CLIP neural network. CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs [31]. Using CLIP, that can be instructed in natural language to predict the most relevant text snippet, given an image, the model has recently merged as a successful representation learner for images. Concretely, CLIP embeddings have several desirable properties: they are robust to image distribution shift, have impressive zero-shot capabilities and have been fine-tuned to achieve state-of-the-art results. In order to obtain a full generative model of images, the CLIP image embedding decoder module is combined with a prior model, which generates possible CLIP image embeddings from a given text caption. We illustrate an image generated from a prompt in Figure 4

IMAGEN : Imagen is a text-to-image diffusion model [22] consisting on large transformer language models. Critically, the main discovery observed with this model made is that large language models, pre-trained on text-only corpora, are very effective at encoding text for image synthesis [34]. Precisely, using Imagen, it has been found out that increasing the size of the language model boosts both sample fidelity and image-text alignment much more than increasing the size of
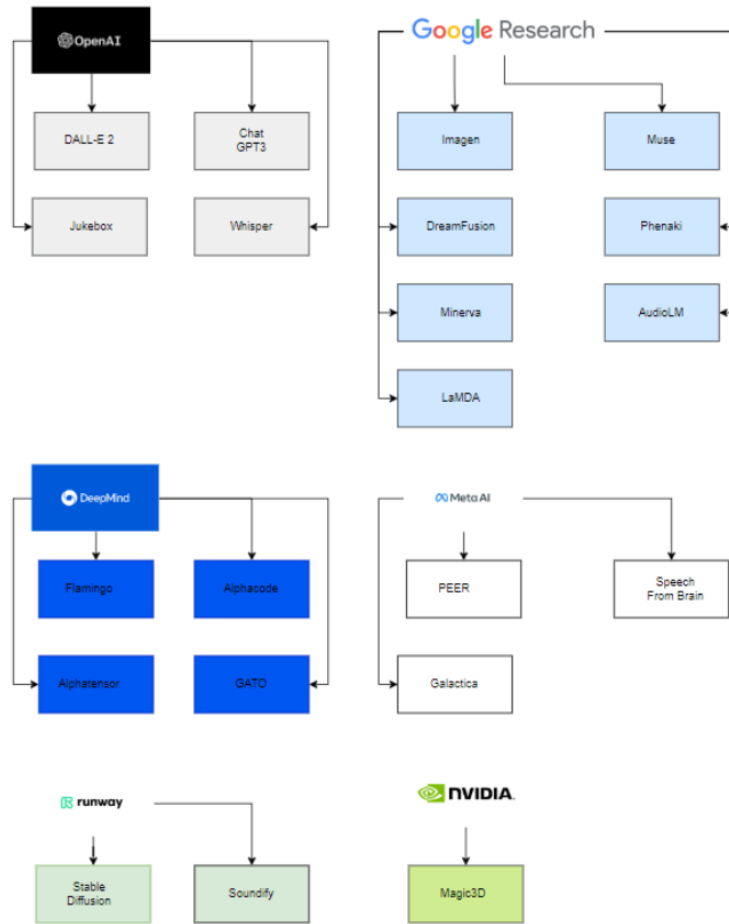
Fig. 3. Models by developer. In terms of major companies participating in startups, note that Microsoft invested 1 billion dollars in OpenAI and helps them with the development of models. As well, note that Google acquired Deepmind in 2014. In terms of universities, note that VisualGPT was developed by KAUST, Carnegie Mellon University and Nanyang Technological University and that the Human Motion Diffusion Model was developed by Tel Aviv University, Israel. As well, other projects are developed by a company in collaboration with a university. Concretely, this is the case for Stable Diffsion (Runway, Stability AI and LMU MUNICH), Soundify (Runway and Carnegie Mellon University) and DreamFusion (Google and UC Berkeley)

the image diffusion model. The model was created by Google and the API can be found in their web page. For the evaluation of their model, Google created Drawbench, a set of 200 prompts that support the evaluation and comparison of text-to-image models. Most concretely, the model is based on a pretrained text encoder (like BERT [13]) that performs a mapping from text to a sequence of

Fig. 4. Image generated from the prompt "A shiba inu wearing a beret and black turtleneck".

word embeddings and a cascade of conditional diffusion models that map these embeddings to images of increasing resolutions. We show an image generated from a prompt in Figure 5.

A cute corgi lives in a house made out of sushi.

Fig. 5. Image generated from the prompt "A cute corgi lives in a house made out of sushi".

Stable Diffusion : Stable Diffusion is a latent-diffusion model that is open-source and has been developed by the CompVis group at LMU Munich. The main difference of this model with respect to the other ones is the use of a latent diffusion model and that it performs image modification as it can perform operations in its latent space. For Stable Diffusion, we can use the API via their website. More concretely, Stable Diffusion consists of two parts: the text encoder and the image generator [17]. The image information creator works completely in the latent space. This property makes it faster than previous diffusion models

that worked in a pixel space. We illustrate a Stable Diffusion image example in Figure 7.

Fig. 6. Image generated from the prompt "A cute corgi lives in a house made out of sushi".

Muse : This model is a Text-to-image transformer model that achieves state-of-the-art image generation while being more efficient than diffusion or autoregressive models [7]. Concretely, it is trained on a masked modelling task in discrete token space. Consequently, it is more efficient because of the use of discrete tokens and requiring fewer sampling iterations. Compared to Parti, a autoregressive model, Muse is more efficient because of parallel decoding. Muse is 10x faster at inference time than Imagen-3B or Parti-3B and 3x faster than Stable Diffusion v 1.4. Muse is also faster than than Stable Diffusion in spite of both models working in the latent space of a VQGAN. We append a comparison of the generated images by DALL·E 2, IMAGEN and Muse in Figure **??**.

3.2 Text-to-3D models

The models that have been described in the previous section deal with the mapping of text prompts to 2D images. However, for some industries like gaming, it is necessary to generate 3D images. In this section, we briefly describe two text-to-3D models: Dreamfusion and Magic3D.

Dreamfusion : DreamFusion is a text-to-3D model developed by Google Research that uses a pretrained 2D text-to-image diffusion model to perform text-to-3D synthesis [30]. In particular, Dreamfusion replaces previous CLIP techniques with a loss derived from distillation of a 2D diffusion model. Concretely, the diffusion model can be used as a loss within a generic continuous optimization problem to generate samples. Critically, sampling in parameter space is much
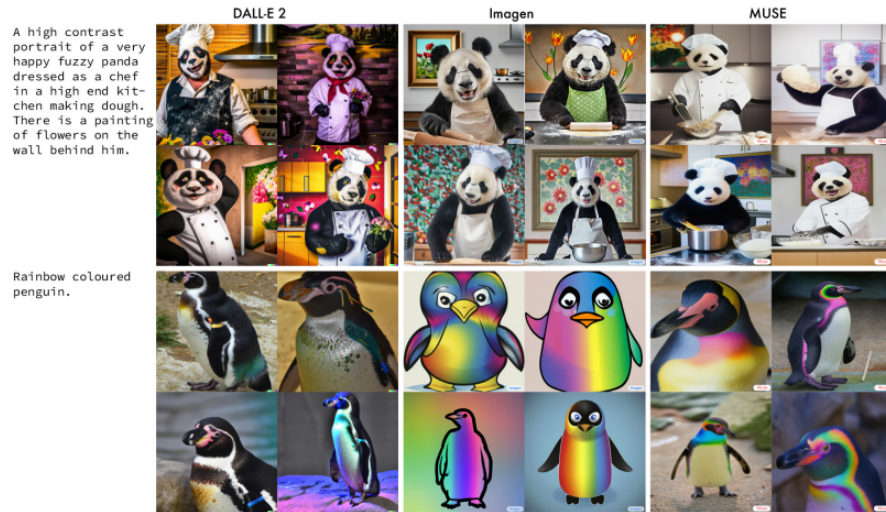
Fig. 7. Comparison of generated images by the DALL·E 2, IMAGEN and Muse models with respect to the prompts that appear in the column of the left. The first column of images contains the results generated by DALL·E 2, the second the results obtained with IMAGEN and the third the images created by Muse.

harder than in pixels as we want to create 3D models that look like good images when rendered from random angles. To solve the issue, this model uses a differentiable generator. Other approaches are focused on sampling pixels, however, this model instead focuses on creating 3D models that look like good images when rendered from random angles. We illustrate in Figure 8 an example of an image created by Dreamfusion from one particular angle along with all the variations that can be generated from additional text prompts. In order to see the full animated image, we recommend to visit the web page of Dreamfusion.

Magic3D : This model is a text to 3D model made by NVIDIA Corporation. While the Dreamfusion model achieves remarkable results, the method has two problems: mainly, the long processing time and the low-quality of the generated images. However, these problems are addressed by Magic3D using a two-stage optimization framework [25]. Firstly, Magic3D builds a low-resolution diffusion prior and, then, it accelerates with a sparse 3D hash grid structure. Using that, a textured 3D mesh model is furthered optimized with an efficient differentiable render. Comparatively, regarding human evaluation, the model achieves better results, as 61.7% prefer this model to DreamFusion. As we can see in Figure 9, Magic3D achieves much higher quality 3D shapes in both geometry and texture compared to DreamFusion.
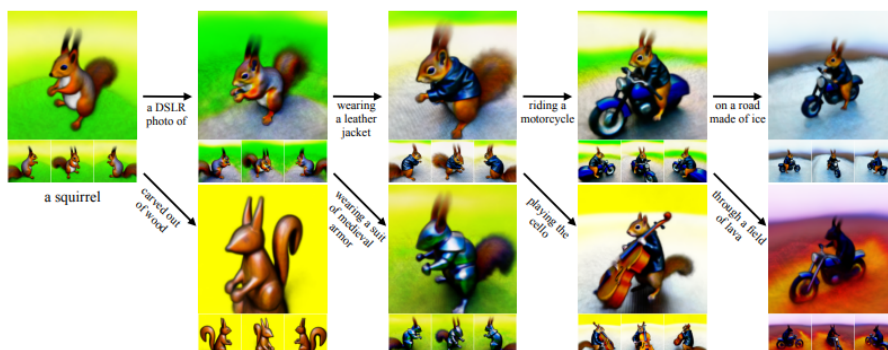
Fig. 8. A 3D squirrel generated by Dreamfusion is shown at the left. Then, the other images contain the modifications generated to the squirrel with text prompts like "wearing a jacket".
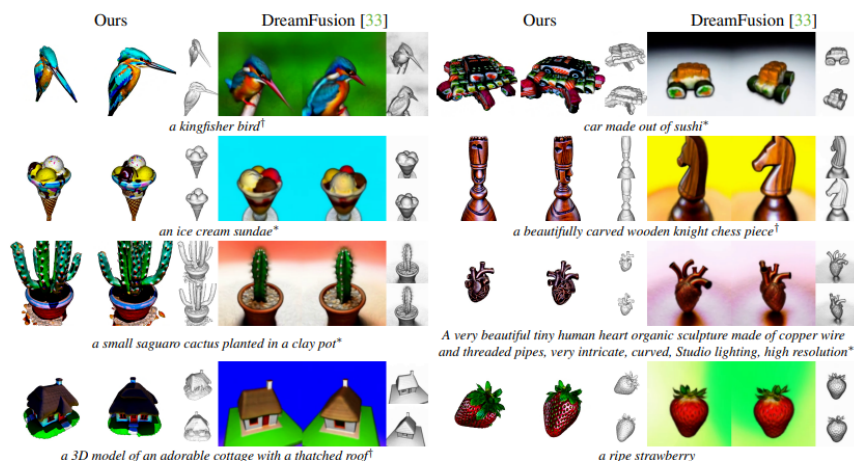

Fig. 9. 3D Images generated by Magic3D and Dreamfusion, where "Ours" refer to Magic3D. We can see a total of 8 text prompts and the images that both models generate from that prompts.

3.3   Image-to-Text models

Sometimes, it is also useful to obtain a text that describes an image, that is precisely the inverse mapping to the one that has been analyzed in the previous subsections. In this section, we analyze two models that perform this task, along with others: Flamingo and VisualGPT.

Flamingo : A Visual Language Model created by Deepmind using few shot learning on a wide range of open-ended vision and language tasks, simply by being prompted with a few input/output examples [2]. Concretely, the input of

Flamingo contains visually conditioned autoregressive text generation models able to ingest a sequence of text tokens interleaved with images and/or videos and produce text as output. A query is made to the model along with a photo or a video and the model answers with a text answer. Some examples can be observed in Figure 10. Flamingo models take advantage of two complementary models: a vision model that analyzes visual scenes and a large language model which performs a basic form of reasoning. The language model is trained on a large amount of text data.
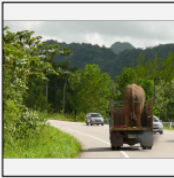

Fig. 10. Input prompts that contain images and text and output generated text responses from Flamingo. Every column contains a single example where we can see how Flamingo answers the question using the image from the text.

VisualGPT : VisualGPT is an image captioning model made by OpenAI [8]. Concretely, VisualGPT leverages knowledge from the pretrained language model GPT-2 [6]. In order to bridge the semantic gap between different modalities, a novel encoder-decoder attention mechanism [39] is designed with an unsaturated rectified gating function. Critically, the biggest advantage of this model is that it does not need for as much data as other image-to-text models. In particular, improving data efficiency in image captioning networks would enable quick data curation, description of rare objects, and applications in specialized domains. Most interestingly, the API of this model can be found on GitHub. We include three examples of text prompts generated by the model with respect to three images fed to the model in Figure 11.

GT: the lady is sitting on the wood bench

Ours	a	woman	sitting	on	a	bench	in	a	park
attention	0.7	0.78	0.82	0.76	0.8	0.96	0.8	0.69	0.85

GT: a laptop with a keyboard and mouse are on this desk

Ours	a	laptop	sitting	on	a	desk	with	a	mouse
attention	0.7	0.78	0.81	0.7	0.7	0.92	0.85	0.64	0.76

GT: a cat is sitting in front of a television

Ours	a	cat	is	sitting	in	front	of	a	television
attention	0.8	0.86	0.8	0.83	0.7	0.72	0.6	0.71	0.93

GT: a number of people sitting on a snowy surface with skis

Ours	a	couple	of	people	sitting	on	a	snowy	surface
attention	0.8	0.87	0.71	0.85	0.91	0.76	0.71	0.94	0.95

Fig. 11. Three examples of text prompts generated by the images shown on the left. We also show the attention scores that the model assign to every word of the texts. In the third image, we can see for example how the most discriminative information about the image is the word "cat" and "television".

3.4   Text-to-Video models

As we have seen in the previous subsections, it is now possible to generate images from text. Consequently, the next logical step is to generate videos, that are sequences of images, from texts. In this section, we provide information about two models that are able to perform this task: Phenaki and Soundify.

Phenaki : This model has been made by Google Research, and it is capable of performing realistic video synthesis, given a sequence of textual prompts [40]. Most interestingly, we can get access to the API of the model from GitHub. In particular, Phenaki is the first model that can generate videos from open domain time variable prompts. To address data issues, it performs joint training on a large image-text pairs dataset as well as a smaller number of video-text exam-

ples can result in generalization beyond what is available in the video datasets. This is mainly due to image-text datasets having billions of inputs while text-video datasets are much smaller. As well, limitations come from computational capabilities for videos of variable length.

The model has three parts: the C-ViViT encoder, the training transformer and the video generator. The encoder gets a compressed representation of videos. First tokens are transformed into embeddings. This is followed by the temporal transformer, then the spatial transformer. After the output of the spatial transformer, they apply a single linear projection without activation to map the tokens back to pixel space. Consequently, the model generates temporally coherent and diverse videos conditioned on open domain prompts even when the prompt is a new composition of concepts. The videos can be minutes long, while the model is trained on 1.4 second videos. Below we show in Figure 12 and in Figure 13 some examples of the creation of a video through a series of text prompts and from a series of text prompts and an image.

Soundify : In video editing, sound in half of the story. But, for professional video editing, the problems come from finding suitable sounds, aligning sounds, video and tuning parameters [26]. In order to solve this issue, Soundify is a system developed by Runway that matches sound effects to video. This system uses quality sound effects libraries and CLIP (a neural network with zero-shot image classification capabilities cited before). Concretely, the system has three parts: classification, synchronization, and mix. The classification matches effects to a video by classifying sound emitters within. To reduce the distinct sound emitters, the video is split based on absolute color histogram distances. In the synchronization part, intervals are identified comparing effects label with each frame and pinpointing consecutive matches above a threshold. In the mix part, effects are split into around one-second chunks. Critically, chunks are stitched via crossfades.

3.5   Text-to-Audio models

As we have seen in the previous subsection, images are not the only important non-structured data format. For videos, for music and in lots of contexts, audio can be critical. Consequently, we analyze in this subsection three models whose input information is text and whose output information is audio.

AudioLM : This model has been made by Google for high-quality audio generation with long-term consistency. In particular, AudioLM maps the input audio into a sequence of discrete tokens and casts audio generation as language modeling task in this representation space [5]. By training on large corpora of raw audio waveforms, AudioLM learns to generate natural and coherent continuations given short prompts. The approach can be extended beyond speech by generating coherent piano music continuations, despite being trained without any symbolic representation of music. As with the other models, the API can
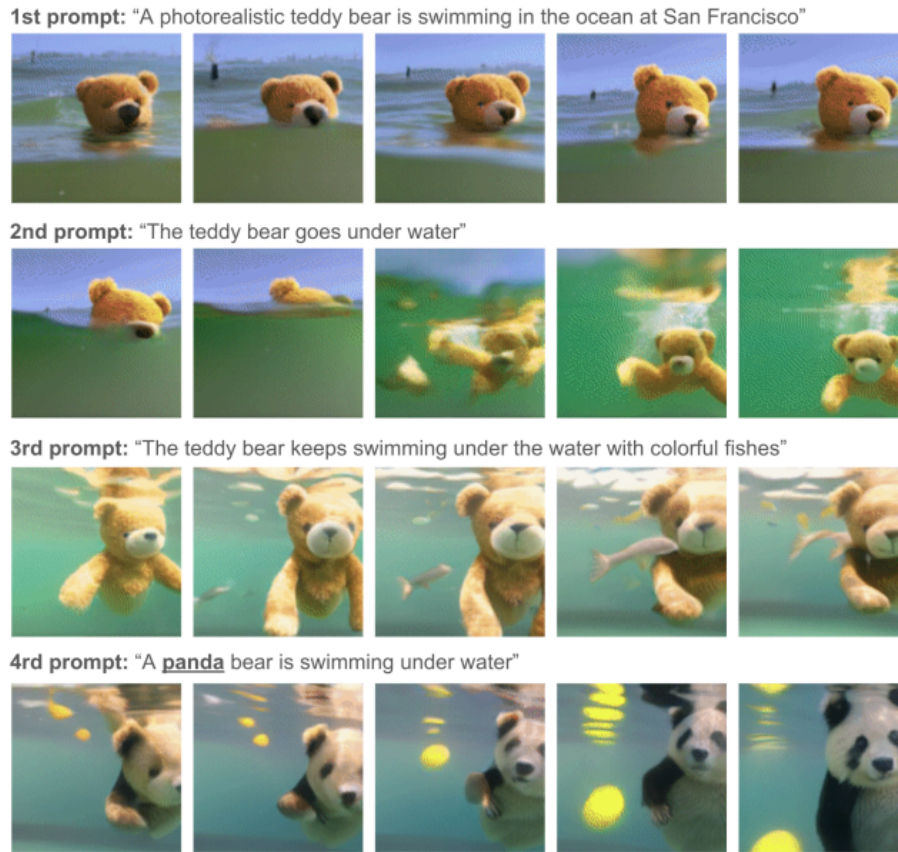
1st prompt: "A photorealistic teddy bear is swimming in the ocean at San Francisco"

2nd prompt: "The teddy bear goes under water"

3rd prompt: "The teddy bear keeps swimming under the water with colorful fishes"

4th prompt: "A **panda** bear is swimming under water"

Fig. 12. Sequence of images created by the Phenaki model given four different prompts.

be found through GitHub. Audio signals involve multiple scales of abstractions. When it comes to audio synthesis, multiple scales make achieving high audio quality while displaying consistency very challenging. This gets achieved by this model by combining recent advances in neural audio compression, self-supervised representation learning and language modelling.

In terms of subjective evaluation, raters were asked to listen to a sample of 10 seconds and decide whether it is human speech or a synthetic continuation. Based on 1000 ratings collected, the rate is 51.2%, which is not statistically significant from assigning labels at random. This tells us that humans cannot differentiate between synthetic and real samples.

Jukebox : This is a model, developed by OpenAI, that generates music with singing in the raw audio domain [14]. Once again, its API can be found in GitHub. Previously, earlier models in the text-to-music genre generated music

Fig. 13. Sequences of images created by the Phenaki model given an image and the prompt. We can see how the model is able to manipulate the given image according to the text prompt.

symbolically in the form of a pianoroll which specifies timing, pitch and velocity. The challenging aspect is the non-symbolic approach where music is tried to be produced directly as a piece of audio. In fact, the space of raw audio is extremely high dimensional which makes the problem very challenging. Consequently, the key issue is that modelling that raw audio produces long-range dependencies, making it computationally challenging to learn the high-level semantics of music.

In order to solve this issue, this model tries to solve it by means of a hierarchical VQ-VAE architecture to compress audio into a discrete space [15], with a loss function designed to retain the most amount of information. This model produces songs from very different genres such as rock, hip-hop and jazz. However, the model is just limited to English songs. Concretely, its dataset for training is from 1.2 million songs from LyricWiki. The VQ-VAE has 5 billion parameters and is trained on 9-second audio clips for 3 days.

Whisper : This model is an Audio-to-Text converter developed by OpenAI. It achieves several tasks in this field: multi-lingual speech recognition, translation and language identification [32]. As in previous cases, its API can be found in the GitHub website. The goal of a speech recognition system should be to work reliably out of the box in a broad range of environments without requiring supervised fine-tuning of a decoder for every deployment distribution. This is hard because of the lack of a high-quality pre-trained decoder.

Concretely, this model is trained on 680,000 hours of labeled audio data. This data is collected from the internet, which results in a very diverse dataset covering a broad distribution of audio from many different environments, record-

ings setups, speakers and languages. The model makes sure that the dataset is only from human voice as machine learning voice would impair the model. Files are broken in 30 second segments paired with the subset of the transcript that occurs within that time segment.

The model has an encoder-deccoder transformer, as this architecture has been validated to scale reliably. We can observe the model's architecture characteristics through the figure below. We can see the different types of data and the learning sequence.

3.6   Text-to-Text models

The previous models all convert a non-structured data type into another one. But, regarding text, it is very useful to convert text into another text in order to satisfy tasks as general question and answering. The following four models treat text and also output texts to satisfy different needs.

ChatGPT : The popular ChatGPT is a model by OpenAI which interacts in a conversational way. As it is widely known, the model answers follow-up questions, challenges incorrect premises and reject inappropriate requests. More concretely, the algorithm behind ChatGPT is based on a transformer. However, the training is made through Reinforcement Learning for Human Feedback. In particular, an initial model is trained using supervised fine-tuning: human AI trainers would provide conversations in which they played both sides, the user and an AI assistant. Then, those people would be given the model-written responses to help them compose their response. This dataset was mixed to that of InstructGPT [4], which was transformed into a dialogue format. A demo can be found in their website and the API may also be found in OpenAI's website. We summarize the main steps of ChatGPT training in Figure 14, available in the ChatGPT demo's website. Finally, ChatGPT is also able to generate code and simple mathematics.

LaMDA : LaMDA is a language model for dialog applications [38]. Unlike most other language models, LaMDA was trained on dialogue. It is a family of transformer-based neural language models specialized for dialog which have up to 137B parameters and are pre-trained on 1.56T words of public dialog data and web text. Fine-tuning can enable for safety and factual grounding of the model. Only 0.001% of training data was used for fine-tuning, which is a great achievement of the model. In particular, dialog modes take advantage of Transformers' ability to present long-term dependencies in text. Concretely, they are generally very well-suited for model scaling. Consequently, LaMDA makes use of a single model to perform multiple tasks: it generates several responses, which are filtered for safety, grounded on an external knowledge source and re-ranked to find the highest-quality response. We illustrate in Figure 15 an example of a dialog with the model.
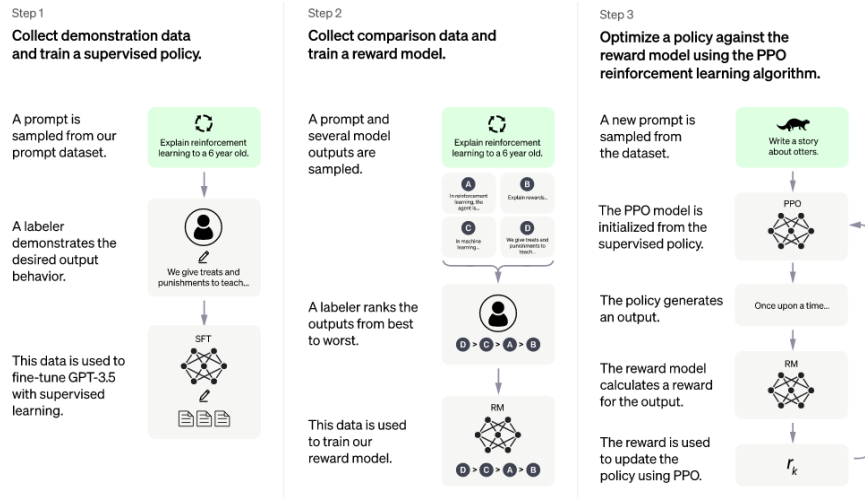
Fig. 14. Training steps of ChatGPT, combining supervised learning with reinforcement learning.

Fig. 15. Example of a dialog made with LaMDA.

PEER : Collaborative language model developed by Meta AI research trained on edit histories to cover the entire writing process [35]. It is based on four steps: Plan, Edit, Explain and Repeat. These steps are repeated until the text is in a satisfactory state that requires no further updates. The model allow to decompose the task of writing a paper into multiple easier subtasks. As well, the model allows humans to intervene at any time and steer the model in any direction.

It is mainly trained on Wikipedia edit histories. The approach is a self-training, using models to infill missing data and then train other models on this synthetic data. The downside of this comes from comments being very noisy

and a lack of citations, which tries to be compensated by a retrieval system which does not always work. The framework is based on an iterative process. The entire process of formulating a plan, collecting documents, performing an edit and explaining it can be repeated multiple times until arriving at a sequence of texts. For the training, a DeepSpeed transformer is used.

Meta AI Speech from Brain : Model developed by Meta AI to help people unable to communicate through speech, typing or gestures [12]. Previous techniques relied on invasive brain-recording techniques which require neurosurgical interventions. This model tries to decode language directly from noninvasive brain recordings. This would provide a safer, more scalable solution that could benefit many more people. The challenge with this proposed method come from noise and differences in each person's brain and where the sensors are placed.

A deep learning model is trained with contrastive learning and used to maximally align noninvasive brain recordings and speech sounds. A self-supervised learning model called wave2vec 2.0. is used to identify the complex representations of speech in the brains of volunteers listening to audiobooks. The two noninvasive technologies used to measure neuronal activity are electroencephalography and magnetoencephalography.

Training data comes from four opensource datasets which represent 150 hours of recordings of 169 volunteers listening to audiobooks. EEG and MEG recordings are inserted into a brain model, which consists of a standard deep convolutional network with residual connections. These recordings are what comes from individuals' brains. This model then has both a speech model for sound and a brain model for MEG data.

Results show that several components of the algorithm were beneficial to decoding performance. As well, analysis shows that the algorithm improves as EEG and MEG recordings increase. This research shows that self-supervised trained AI can decode perveived speech despite noise and variability in that data. The biggest limitation of this research is that it focuses on speech perception, but the ultimate goal would be to extend this work to speech production.

3.7   Text-to-Code models

Although we have covered text-to-text models, not all texts follows the same syntax. An special type of text is code. In programming, it is essential to know how to convert an idea into code. In order to do so, Codex and Alphacode models help.

Codex : AI system created by OpenAI which translates text to code. It is a general-purpose programming model, as it can be applied to basically any programming task [9]. Programming can be broken down into two parts: breaking a problem down into simpler problems and mapping those problems into existing code (libraries, APIs, or functions) that already exist. The second part is the most time-barring part for programmers, and it is where Codex excels the most.

The data collected for training was collected in May 2020 from public software repositories hosted on GitHub, containing 179GB of unique Python files under 1 MB. The model is fine-tuned from GPT-3, which already contains strong natural language representations. The demo and the API can be found in Open AI's website.

Alphacode : Other language models have demonstrated an impressive ability to generate code, but these systems still perform poorly when evaluated on more complex, unseen problems. However, Alphacode is a system for code generation for problems that require for deeper reasoning [24]. Three components are key for this achievement: having an extensive dataset for training and evaluation, large and efficient transformer based architectures and a large-scale model sampling.

In terms of training, the model is firstly pre-trained through GitHub repositories amounting to 715.1 GB of code. This is a much more extensive dataset than Codex's pre training dataset. For the training to be better, a fine-tuning dataset is introduced from the Codeforces plataform. Through this platform, Codecontests are conducted, for the validation phase, in which we better the performace of the model. Regarding the transformer-based architecture, they use an encoder-decoder transformer architecture. Compared to decoder-only architectures commonly used, this architecture allows for a bidirectional description and extra flexibility. As well, they use a shallow encoder and a deep encoder to further the model's efficiency. To reduce the cost of sampling, multi-query attention is used.

3.8   Text-to-Science models

Even scientific texts are being targeted by generative AI, as the Galactica and Minerva models have shown. Although there is a long way to manage success in this field, it is critical to study the first attempts towards automatic scientific text generation.

Galactica : Galactica is a new large model for automatically organizing science developed by Meta AI and Papers with Code. The main advantage of the model is the ability to train on it for multiple epochs without overfitting, where upstream and downstream performance improves with use of repeated tokens. The dataset design is critical to the approach as all of it is processed in a common markdown format to blend knowledge between sources. Citations are processed via a certain token that allows researchers to predict a citation given any input context. The capability of the model of predicting citations improves with scale and the model becomes better at the distribution of citations. In addition, the model can perform multi-modal tasks involving SMILES chemical formulas and protein sequences. Concretely, Galactica uses a transformer architecture in a decoder-only setup with GeLU activation for all model sizes.

Minerva : Language model capable of solving mathematical and scientific questions using step-by-step reasoning. Minerva has a very clear focus on the collection of training data for this purpose. It solves quantitative reasoning problems, makes models at scale and employs best-in-class inference techniques. Concretely, Minerva solves these problems by generating solutions step-by-step, this means including calculations and symbolic manipulation without having the need for external tools such a calculator.

3.9   Other models

We would like to finish our review by covering additional models that do not fit any of the categories mentioned previously.

Alphatensor, created by the research company Deepmind, is a completely revolutionary model in the industry for its ability to discover new algorithms [16]. In the published example, Alpha Tensor creates a more efficient algorithm for matrix multiplication, which is very important, as improving the efficiency of algorithms affects a lot of computations, from neural networks to scientific computing routines.

The methodology is based on a deep reinforcement learning approach in which the agent, AlphaTensor is trained to play a single-player game where the objective is finding tensor decomposisitions within a finite factor space. At each step of the TensorGame, the player selects how to combine different entries of the matrices to multiply. A score is assigned based on the number of selected operations required to reach the correct multiplication result. To solve TensorGame, an agent, AlphaTensor was developed. AlphaTensor uses a specialized neural network architecture to exploit symmetries using synthetic training games.

GATO is a single generalist agent made by Deepmind. It works as a multimodal, multi-task, multi-embodiment generalist policy [33]. The same network with the same weights can carry very different capabilities from playing Atari, caption images, chatting, stacking blocks and many more. There are many benefits from using a single neural sequence model across all tasks. It reduces the need for hand crafting policy models with their own inductive biases. It increases the amount and diversity of training data. This general agent is successful at many tasks and can be adapted with little extra data to succeed at an even larger number of tasks. r training at the operating point of model scale that allows real-time control of real-world robots, currently around 1.2B parameters in the case of GATO.

Other published generative AI models are able to generate human motion [37] or, in the case of ChatBCG, slides using ChatGPT as a surrogate model.

4   Ethical issues of Generative AI models

Most AI Ethics guidelines converge around five topics: transparency, justice, non-maleficence, responsibility, and privacy [20]. In terms of transparency and justice, the analyzed generative AI models generally use transformers, which are

very hard to understand when conducting audits by data protection offices as they are complex black-box models [23]. Critically, as interpretability is key at the moment in order to make correct regulation, research must be conducted to make these models more interpretable.

Regarding non-maleficence, we can observe the example of OPENAI, requiring a large labor force in Kenya to look at harmful situations such as suicide, incest or sexual abuse to purge it out of the model [1]. While purging these situations out of the model may seem ethical, the way in which OpenAI is achieving that is controversial.

Regarding privacy, some ethical experts have alluded to data minimization. This principle means that we should only collect personal information to what is directly relevant and necessary [17]. This can be very hard when we need to look at such large datasets as the ones we are looking at in Generative AI. As an example, the LaMDA model is pre-trained on a dataset of 1.56T words, thus making the principle of data minimization hard.

Finally, it is critical to clearly state that these generative AI models are, in essence, statistical information processors over enormous datasets whose software instructions are processed in a computer. As their behaviour is just a gigantic set of binary instructions executed on a CPU or GPU, these models, although they exhibit behaviour that may seem unique to humans, are not humans nor we can attribute to them awareness just because of the computational intelligence that they exhibit [28, 18]. For example, these models lack of an understanding of the problems that they solve nor they perceive the qualia associated to the problem [19]. Consequently, they must be a tool, or an assistant, for human experts to decide about ethical decisions, but they should not decide themselves alone.

5   Conclusions and further work

Through this paper, we can observe the capabilities which generative artificial intelligence has. We have seen a great deal of creativity as well as personalization in tasks such as text-to-image or in tasks such as text-to-audio. They also are accurate in text-to-science or text-to-code tasks. This can help economies in a major way as it can help optimize creative and non-creative tasks.

However, because of the way that they are constructed at the moment, these models face a number of limitations. In terms of dataset, finding data for some of the models found such as the text-to-science or the text-to-audio is very hard, making it very time-consuming to train the model. In particular, datasets and parameters have to be enormous, making it harder to train. One of the biggest issues with models is trying solutions out of the problems in the dataset, with which models have more trouble solving. As well, in terms of computation, a lot of time and computation capacity is necessary in order to run them. Many days and advanced computers are needed in order to run the models.

In addition, these models face bias from the data which needs to be controlled. Galactica model tries to control this issue through a layer of no bias, but it still a major issue for Generative Artificial Intelligence.

With the Minerva model, we can see that the model knows the steps which it needs to take to solve an equation. This is groundbreaking as one of the biggest limitations with these models is that the models do not understand exactly what they are doing. Moreover, it's still an industry starting; thus accuracy is still an issue. Text-to-video models for example are only represented by Phenaki because how hard it is to produce accurate videos. Text-to-science models do find some accuracy but that accuracy is still way behind to what it should be for professionals to actually rely on this technology on a day-to-day basis.

Furthermore, these models need to be constrained because of a lack of understanding of ethics. Phenaki on its paper even acknowledges that a system like text-to-video can be used to create deep-fakes. Lastly, we are still in a phase where we are discovering what exactly the purpose of this intelligence will be. There has been articles comparing Google to ChatGPT, which is totally inexact as ChatGPT does not update its information in real time. We should be aware about the limitations of these models to try and improve them in the following years.

References

1. OpenAI used kenyan workers on less than $2 per hour to make chatGPT less toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/, 2023.

2. ALAYRAC, J.-B., DONAHUE, J., LUC, P., MIECH, A., BARR, I., HASSON, Y., LENC, K., MENSCH, A., MILLICAN, K., REYNOLDS, M., ET AL. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

3. ANANTRASIRICHAI, N., AND BULL, D. Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review* (2021), 1–68.

4. BHAVYA, B., XIONG, J., AND ZHAI, C. Analogy generation by prompting large language models: A case study of instructgpt. *arXiv preprint arXiv:2210.04186* (2022).

5. BORSOS, Z., MARINIER, R., VINCENT, D., KHARITONOV, E., PIETQUIN, O., SHARIFI, M., TEBOUL, O., GRANGIER, D., TAGLIASACCHI, M., AND ZEGHIDOUR, N. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143* (2022).

6. BUDZIANOWSKI, P., AND VULIĆ, 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).

7. CHANG, H., ZHANG, H., BARBER, J., MASCHINOT, A., LEZAMA, J., JIANG, L., YANG, M.-H., MURPHY, K., FREEMAN, W. T., RUBINSTEIN, M., ET AL. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

8. CHEN, J., GUO, H., YI, K., LI, B., AND ELHOSEINY, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18030–18040.

9. CHEN, M., TWOREK, J., JUN, H., YUAN, Q., PINTO, H. P. D. O., KAPLAN, J., EDWARDS, H., BURDA, Y., JOSEPH, N., BROCKMAN, G., ET AL. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

10. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine 35*, 1 (2018), 53–65.
11. Daras, G., and Dimakis, A. G. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169* (2022).
12. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266* (2022).
13. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
14. Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).
15. Ding, S., and Gutierrez-Osuna, R. Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion. In *INTERSPEECH* (2019), pp. 724–728.
16. Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature 610*, 7930 (2022), 47–53.
17. Galdon Clavell, G., Martín Zamorano, M., Castillo, C., Smith, O., and Matic, A. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), pp. 265–271.
18. Garrido-Merchán, E. C., Molina, M., and Mendoza-Soto, F. M. A global workspace model implementation and its relations with philosophy of mind. *Journal of Artificial Intelligence and Consciousness 9*, 01 (2022), 1–28.
19. Garrido-Merchán, E. C., and Blanco, C. Do artificial intelligence systems understand? *arXiv preprint arXiv:2207.11089* (2022).
20. Jobin, A., Ienca, M., and Vayena, E. The global landscape of ai ethics guidelines. *Nature Machine Intelligence 1*, 9 (2019), 389–399.
21. Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., and Huber, P. Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE Frontiers in Education Conference (FIE)* (2016), IEEE, pp. 1–9.
22. Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems 34* (2021), 21696–21707.
23. LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436–444.
24. Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science 378*, 6624 (2022), 1092–1097.
25. Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440* (2022).
26. Lin, D. C.-E., Germanidis, A., Valenzuela, C., Shi, Y., and Martelaro, N. Soundify: Matching sound effects to video. *arXiv preprint arXiv:2112.09726* (2021).
27. Lin, T., Wang, Y., Liu, X., and Qiu, X. A survey of transformers. *AI Open* (2022).

28. MERCHÁN, E. C. G., AND LUMBRERAS, S. On the independence between phenomenal consciousness and computational intelligence. *arXiv preprint arXiv:2208.02187* (2022).

29. MURPHY, K. P. *Probabilistic machine learning: an introduction.* MIT press, 2022.

30. POOLE, B., JAIN, A., BARRON, J. T., AND MILDENHALL, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

31. RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., ET AL. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763.

32. RADFORD, A., KIM, J. W., XU, T., BROCKMAN, G., MCLEAVEY, C., AND SUTSKEVER, I. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

33. REED, S., ZOLNA, K., PARISOTTO, E., COLMENAREJO, S. G., NOVIKOV, A., BARTH-MARON, G., GIMENEZ, M., SULSKY, Y., KAY, J., SPRINGENBERG, J. T., ET AL. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

34. SAHARIA, C., CHAN, W., SAXENA, S., LI, L., WHANG, J., DENTON, E., GHASEMIPOUR, S. K. S., AYAN, B. K., MAHDAVI, S. S., LOPES, R. G., ET AL. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487* (2022).

35. SCHICK, T., DWIVEDI-YU, J., JIANG, Z., PETRONI, F., LEWIS, P., IZACARD, G., YOU, Q., NALMPANTIS, C., GRAVE, E., AND RIEDEL, S. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663* (2022).

36. SUSNJAK, T. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292* (2022).

37. TEVET, G., RAAB, S., GORDON, B., SHAFIR, Y., COHEN-OR, D., AND BERMANO, A. H. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

38. THOPPILAN, R., DE FREITAS, D., HALL, J., SHAZEER, N., KULSHRESHTHA, A., CHENG, H.-T., JIN, A., BOS, T., BAKER, L., DU, Y., ET AL. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

39. VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

40. VILLEGAS, R., BABAEIZADEH, M., KINDERMANS, P.-J., MORALDO, H., ZHANG, H., SAFFAR, M. T., CASTRO, S., KUNZE, J., AND ERHAN, D. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022).