REVA UNIVERSITY
Established as per the Section 2(f) of the UGC Act, 1956
Approved by AICTE, COA and BCI, New Delhi

Generative AI Models

Dr. Nimrita Koul
School of CSE, REVA University
https://www.linkedin.com/in/nimritakoul/

Outline

- Introduction to Generative AI
- Deep Neural Network Architectures for GenAI-
  - GANS, VAE, Stable Diffusion, Transformers
- Training Large Language Models
- Prompting
- Tools, Use Cases
- Ethical Use of Gen AI
- Challenges
- Research Directions

Objectives and Outcomes

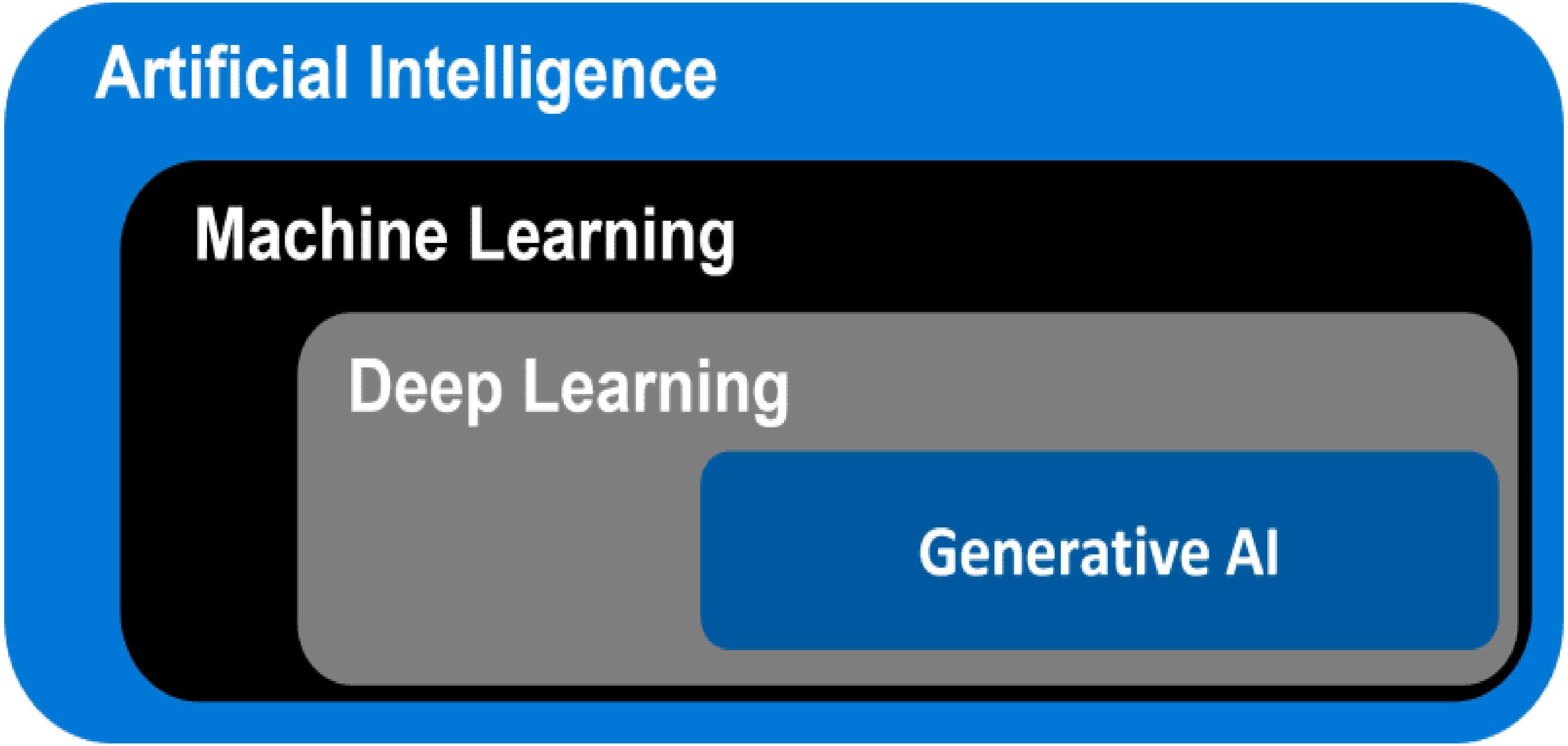
Objectives:

1. To give a concise introduction to generative AI models and their workings
2. Explain models Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models like Transformers.
3. Discuss the challenges and limitations.

Outcomes:

1. By the end of this talk, the audience should understand the fundamental concepts and workings of generative AI models.
2. Audience should appreciate the diverse applications of Gen AI.
3. I should be able to pique your interest in exploring generative AI further and using it in your own projects.

Generative Artificial Intelligence (GenAI)

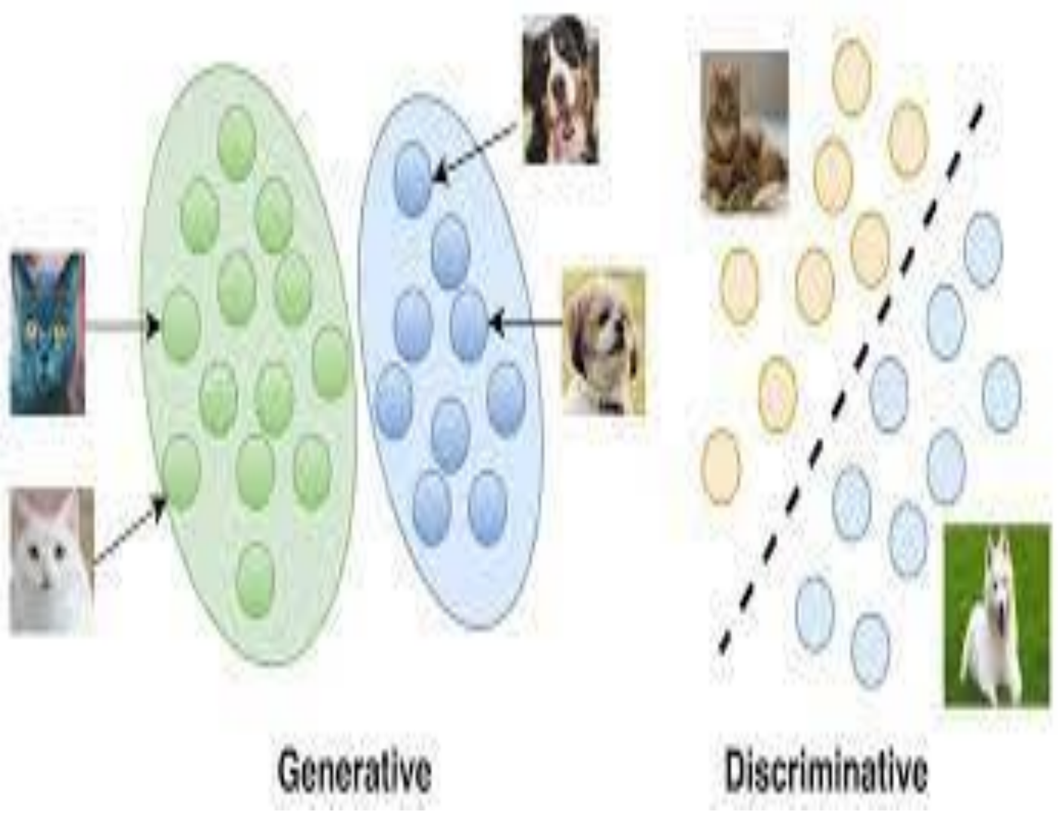


Source: https://dev.to/mitul3737/artificial-intelligence-vs-machine-learning-vs-deep-learning-vs-generative-ai-7ae

Discriminative and Generative Deep Learning Models

- Discriminative Models or conditional models are used for supervised learning. E.g., Logistic Regression, SVM, Decision Trees, Neural Nets, KNN, Random Forest etc.

- Generative Models can be used to generate new data points (Text, images, audio, video, animation, 3D models etc.) that are like the data used for training these models. These usually are applied to solve unsupervised ML tasks. E.g., Bayesian networks, autoregressive models, generative adversarial networks, HMMS, LDA etc.
- Generative language models such as GPT, LaMDA, PaLM, Llama, learn about patterns in language through training data and given some text, they predict what comes next.



Source: https://vitalflux.com/generative-vs-discriminative-models-examples/

The absolute basics - Bayes Classifier

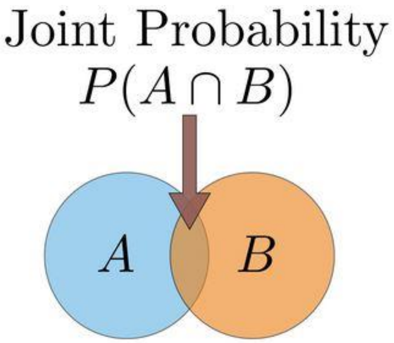
- Let's say you have input data x and you want to classify the data into labels y. You can do this using Naïve Bayes Classifier or Logistic Regression classifier. Bayes is a generative model, LR is discriminative.
- A generative model learns the joint probability distribution p(x,y) and a discriminative model learns the conditional probability distribution p(y|x).
- E.g., we have data in the form (x,y): (1,0), (1,0), (2,0), (2, 1)

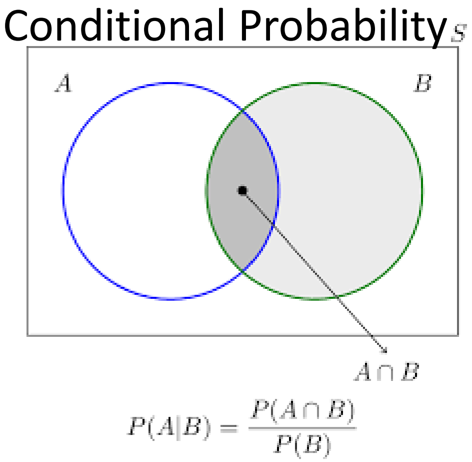
`p(x,y)` is

	y=0	y=1
x=1	1/2	0
x=2	1/4	1/4

`p(y|x)` is

	y=0	y=1
x=1	1	0
x=2	1/2	1/2





Source: https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-a-discriminative-algorithm

Generative AI

- **Generative Artificial Intelligence (GenAI)** models **create new content based** on their training from the existing content.
- When given a prompt, GenAI uses a statistical model learnt during training to predict the response. This prediction is called as **Completion.**
- **Large Language Models** (LLMs) are Gen AI models that generate text content. They learn representation of language based on patterns in training data, then given a prompt, they predict the next word in sequence.
- **Generative Image Models** learn to create new images using techniques like diffusion. Given a prompt, they transform random noise into images.
- E.g., Image generators (such as Midjourney or Stable Diffusion), large language models (such as GPT-4, PaLM, or Claude), code generation tools (such as Copilot), or audio generation tools (such as VALL-E or resemble.ai).

Generative AI's evolution

For an advanced technology that's considered relatively new, generative AI is deep-rooted in history and innovation.

1932
Georges Artsrouni invents a machine he reportedly called the **"mechanical brain"** to translate between languages on a mechanical computer encoded onto punch cards.

1966
MIT professor Joseph Weizenbaum creates the first chatbot, **Eliza**, which simulates conversations with a psychotherapist.

1980
Michael Toy and Glenn Wichman develop the Unix-based game *Rogue*, which uses procedural content generation to dynamically generate new game levels.

1986
Michael Irwin Jordan lays the foundation for the modern use of recurrent neural networks (RNNs) with the publication of "Serial order: a parallel distributed processing approach."

2000
University of Montreal researchers publish "A Neural Probabilistic Language Model," which suggests a method to model language using feed-forward neural networks.

2011
Apple releases **Siri**, a voice-powered personal assistant that can generate responses and take actions in response to voice requests.

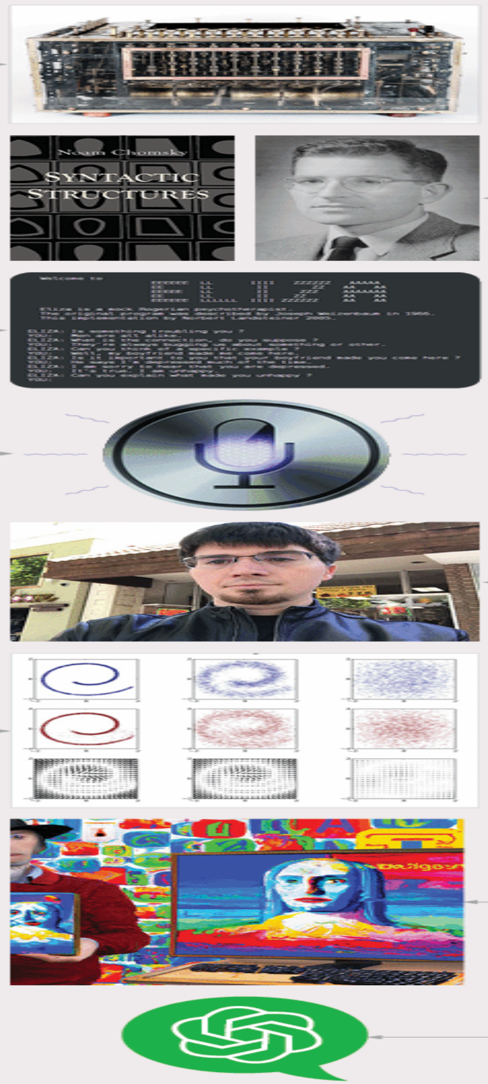
2013
Google researcher Tomas Mikolov and colleagues introduce word2vec to identify semantic relationships between words automatically.

2015
Stanford researchers publish work on diffusion models in the paper **"Deep Unsupervised Learning using Nonequilibrium Thermodynamics."** The technique provides a way to reverse-engineer the process of adding noise to a final image.

2018
Google researchers implement transformers into BERT, which is trained on more than 3.3 billion words and can automatically learn the relationship between words in sentences, paragraphs and even books to predict the meaning of text. It has 110 million parameters.

Google DeepMind researchers develop AlphaFold for predicting protein structures, laying the foundation for generative AI applications in medical research, drug development and chemistry.

OpenAI releases GPT (Generative Pre-trained Transformer). Trained on about 40 gigabytes of data and consisting of 117 million parameters, GPT paves the way for subsequent LLMs in content generation, chatbots and language translation.



1957
Linguist **Noam Chomsky** publishes ***Syntactic Structures***, which describes grammatical rules for parsing and generating natural language sentences.

1968
Computer science professor Terry Winograd creates SHRDLU, the first multimodal AI that can manipulate and reason out a world of blocks according to instructions from a user.

1985
Computer scientist and philosopher Judea Pearl introduces Bayesian networks causal analysis, which provides statistical techniques for representing uncertainty that leads to methods for generating content in a specific style, tone or length.

1989
Yann LeCun, Yoshua Bengio and Patrick Haffner demonstrate how convolutional neural networks (CNNs) can be used to recognize images.

2006
Data scientist Fei-Fei Li sets up the ImageNet database, which provides the foundation for visual object recognition.

2012
Alex Krizhevsky designs the AlexNet CNN architecture, pioneering a new way of automatically training neural networks that take advantage of recent GPU advances.

2014
Research scientist **Ian Goodfellow** develops generative adversarial networks (GANs), which pit two neural networks against each other to generate increasingly realistic content.

Diederik Kingma and Max Welling introduce variational autoencoders to generate images, videos and text.

2017
Google researchers develop the concept of transformers in the seminal paper "Attention is all you need," inspiring subsequent research into tools that could automatically parse unlabeled text into large language models (LLMs).

2021
OpenAI introduces **Dall-E**, which can generate images from text prompts. The name is a combination of WALL-E, the name of a fictional robot, and the artist Salvador Dali.

2022
Researchers from Runway Research, Stability AI and CompVis LMU release Stable Diffusion as open source code that can automatically generate image content from a text prompt.

OpenAI releases **ChatGPT** in November to provide a chat-based interface to its GPT 3.5 LLM. It attracts over 100 million users within two months, representing the fastest ever consumer adoption of a service.

2023
Getty Images and a group of artists separately sue several companies that implemented Stable Diffusion for copyright infringement.

Microsoft integrates a version of ChatGPT into its Bing search engine. Google quickly follows with plans to release the Bard chat service based on its Lamda engine. And the controversy over detecting AI-generated content heats up.

Source: https://cdn.ttgtmedia.com/rms/onlineimages/generative_ai_evolution_desktop.png

Date	Name	Organization	Language	# Parameters	# Tokens	Architecture	Train. Obj.	Tokenizer	Pos. Embed.	IF	MoE	Code avail.	Ckpt. avail.	Pre-trained
2018.11	GPipe [226]	Google	Multil.	6B	N/A	Enc. & Dec.	NTP	BPE	Learned	✗	✗	✓	✗	✗
2019.09	Megatron-LM [501]	Microsoft	Eng.	8.3B	157B	Dec.-Only	NTP	BPE	Learned	✗	✗	✓	✗	✗
2019.10	T5 [443]	Google	Multil.	11B	1T	Enc. & Dec.	SC	SP	T5	✗	✗	✓	✓	✗
2020.05	GPT-3 [59]	OpenAI	Eng.	175B	300B	Dec.-Only	NTP	BPE	Learned	✗	✗	✗	✗	✗
2020.06	GShard [298]	Google	Multil.	600B	1T	Enc. & Dec.	NTP	SP	N/A	✗	✓	✗	✗	✗
2020.10	mT5 [631]	Google	Multil.	13B	1T	Enc. & Dec.	SC	SP	T5	✗	✗	✓	✓	✗
2021.01	Switch [145]	Google	Multil.	1.5T	N/A	Enc. & Dec.	SC	SP	T5	✗	✓	✓	✓	✗
2021.03	BASE [302]	Meta	Eng.	117B	N/A	Enc. & Dec.	NTP	BPE	Sinus.	✗	✓	✓	✗	✗
2021.04	PanGu-α [659]	Huawei	Multil.	200B	317B	Dec.-Only	NTP	BPE	Learned	✗	✗	✗	✗	✗
2021.05	ByT5 [630]	Google	Multil.	12.9B	1T	Enc. & Dec.	SC	N/A	T5	✗	✗	✓	✓	✗
2021.06	CPM-2 [669]	Tsinghua Uni.	Multil.	198B	N/A	Enc. & Dec.	SC	Custom	Sinus.	✗	✓	✓	✓	✗
2021.06	nmT5 [255]	Google	Multil.	3.7B	100B	Enc. & Dec.	MLM, NTP	SP	T5	✗	✗	✗	✗	✓
2021.07	ERNIE 3.0 [530]	Baidu	Chin.	10B	375B	Enc. & Dec.	Custom	BPE	Rel.	✗	✗	✗	✗	✗
2021.08	Jurassic-1 [319]	AI21	Eng.	178B	300B	Enc. & Dec.	NTP	SP	Learned	✗	✗	✗	✗	✗
2021.08	ExT5 [23]	Google	Eng.	11B	1T	Enc. & Dec.	SC, Custom	SP	T5	✗	✗	✓	✗	✗
2022.01	FLAN-LaMDA [598]	Google	Eng.	137B	245M	Dec.-Only	NTP	BPE	T5	✗	✓	✗	✗	✓
2021.10	M6-10T [322]	Alibaba	Eng.	10T	N/A	Uni. Enc. & Dec.	SC, NTP	SP	N/A	✗	✗	✗	✗	✗
2021.10	Yuan [615]	Inspur AI	Chin.	245B	180B	Dec.-Only	NTP	BPE	N/A	✗	✗	✗	✗	✗
2021.10	T0 [475]	BigScience	Eng.	11B	12B	Enc. & Dec.	SC, NTP	SP	T5	✗	✗	✓	✓	✓
2021.12	Gopher [441]	DeepMind	Eng.	280B	300B	Dec.-Only	NTP	SP	Rel.	✗	✗	✗	✗	✗
2021.12	RETRO [52]	DeepMind	Eng.	7B	419B	Enc. & Dec.	NTP (Ret.)	SP	Rel.	✗	✗	✗	✗	✗
2021.12	GLaM [130]	Google	Multil.	1.2T	600B	Dec.-Only	NTP	SP	Rel.	✗	✓	✗	✗	✗
2021.12	WebGPT [379]	OpenAI	Eng.	175B	N/A	Dec.-Only	NTP	BPE	Learned	✗	✗	✗	✗	✓
2021.12	FairSeq [400]	Meta	Eng.	1.1T	300B	Dec.-Only	NTP	BPE	Sinus.	✗	✓	✓	✓	✗
2021.12	XGLM [324]	Meta	Multil.	7.5B	500B	Dec.-Only	NTP	Unigram	Sinus.	✗	✗	✓	✓	✗
2022.01	LaMDA [551]	Google	Eng.	137B	768B	Dec.-Only	NTP	BPE	T5	✗	✗	✗	✗	✗
2022.01	MT-NLG [515]	Microsoft	Eng.	530B	270B	Dec.-Only	NTP	BPE	Sinus.	✗	✗	✗	✗	✗
2022.02	ST-MoE [687]	Google	Eng.	269B	1.5T	Enc. & Dec.	SC	SP	Sinus.	✗	✓	✓	✗	✗
2022.03	InstructGPT [403]	OpenAI	Eng.	175B	N/A	Dec.-Only	RLHF	BPE	Learned	✓	✗	✗	✗	✓
2022.03	GopherCite [362]	DeepMind	Eng.	280B	N/A	Dec.-Only	RLHF	BPE	Rel.	✓	✗	✗	✗	✓
2022.03	sMLP [653]	Meta	Eng.	9.4B	N/A	Enc. & Dec.	NTP	BPE	Sinus.	✗	✓	✗	✗	✗
2022.03	Chinchilla [206]	DeepMind	Eng.	70B	1.4T	Dec.-Only	NTP	SP	Rel.	✗	✗	✗	✗	✗
2022.04	PaLM [86]	Google	Multil.	540B	780B	Dec.-Only	NTP	SP	RoPE	✗	✓	✗	✗	✗
2022.04	GPT-NeoX [47]	EleutherAI	Eng.	20B	472B	Dec.-Only	NTP	BPE	RoPE	✗	✗	✓	✓	✗
2022.04	Tk-Instruct [589]	AI2	Eng.	11B	1B	Enc. & Dec.	NTP	SP	T5	✓	✗	✓	✓	✗
2022.04	METRO-LM [33]	Microsoft	Eng.	5.4B	2T	Enc.-Only	METRO	SP	T5	✗	✗	✗	✗	✗
2022.04	mGPT [500]	Sber	Multi.	13B	440B	Dec.-Only	NTP	BPE	Learned	✗	✗	✓	✓	✗
2022.05	OPT [666]	Meta	Eng.	175B	300B	Dec.-Only	NTP	BPE	Learned	✗	✗	✓	✓	✗
2022.05	UL2 [545]	Google	Eng.	20B	1T	Enc. & Dec.	MoD	Unigram	T5	✗	✗	✗	✓	✗
2022.05	DeepStruct [578]	UC Berkeley	Eng.	10B	N/A	Enc. & Dec.	Struc.	BPE	Sinus.	✗	✗	✗	✗	✗
2022.07	Minerva [305]	Google	Eng.	540B	26B	Dec.-Only	NTP	SP	RoPE	✗	✗	✗	✗	✗
2022.08	PEER [482]	Meta	Eng.	11B	5B	Enc. & Dec.	NTP	SP	T5	✗	✗	✗	✗	✓
2022.08	AlexaTM [517]	Amazon	Multil.	20B	1T	Enc. & Dec.	MoD, NTP	SP	Sinus.	✗	✗	✗	✓	✓
2022.10	GLM-130B [658]	Tsinghua Uni.	Multil.	130B	400B	Uni. Enc. & Dec.	ARBF	SP	RoPE	✗	✗	✓	✓	✗
2022.10	U-PaLM [547]	Google	Eng.	540B	1.3B	Dec.-Only	MoD	SP	RoPE	✗	✓	✗	✗	✓
2022.10	FLAN-PaLM [93]	Google	Eng.	540B	1.4B	Dec.-Only	NTP	SP	RoPE	✓	✓	✗	✗	✓
2022.11	BLOOM [479]	BigScience	Multil.	176B	366B	Dec.-Only	NTP	BPE	ALiBi	✗	✗	✓	✓	✗
2022.11	Galactica [548]	Meta	Eng.	120B	450B	Dec.-Only	NTP	BPE	Learned	✗	✗	✓	✓	✗
2022.11	Atlas [236]	Meta	Eng.	11B	N/A	Enc. & Dec.	MLM	BPE	T5	✗	✗	✓	✓	✓
2022.11	BLOOMZ [377]	BigScience	Multil.	176B	13B	Dec.-Only	NTP	BPE	ALiBi	✓	✗	✓	✓	✓
2022.11	mT0 [377]	BigScience	Multil.	13B	13B	Enc. & Dec.	NTP	SP	T5	✓	✗	✓	✓	✓
2022.12	OPT-IML [235]	Meta	Eng.	175B	2B	Dec.-Only	NTP	BPE	Sinus.	✓	✗	✓	✓	✓
2022.12	Med-PaLM [511]	Google	Eng.	540B	0B	Dec.-Only	NTP	SP	RoPE	✗	✗	✗	✗	✓
2023.02	LLaMA{-I} [556]	Meta	Eng.	65B	1.4T	Dec.-Only	NTP	BPE	RoPE	✓	✗	✓	✓	✗
2023.03	PanGu-Σ [455]	Huawei	Multil.	1T	329B	Dec.-Only	NTP	BPE	Learned	✗	✓	✗	✗	✓
2023.03	CoLT5 [15]	Google	Eng.	5.3B	1T	Enc. & Dec.	MoD	N/A	T5	✗	✗	✗	✗	✗
2023.03	BloombergGPT [616]	Bloomberg	Eng.	50B	569B	Dec.-Only	NTP	Unigram	ALiBi	✗	✗	✗	✗	✗
2023.04	Cerebras-GPT [121]	Cerebras	Eng.	13B	257B	Dec.-Only	NTP	BPE	RoPE	✗	✗	✗	✓	✗
2023.04	Pythia [44]	EleutherAI	Eng.	12B	300B	Dec.-Only	NTP	BPE	RoPE	✗	✗	✓	✓	✗
2023.04	WizardLM [625]	Microsoft	Eng.	30B	N/A	Dec.-Only	NTP	BPE	RoPE	✓	✗	✓	✓	✓
2023.05	Guanaco [118]	Univ. of Washington	Multil.	65B	82M	Dec.-Only	NTP	BPE	RoPE	✓	✗	✗	✓	✓
2023.04	RWKV [417]	RWKV	Eng.	14B	N/A	Dec.-Only	NTP	BPE	RoPE	✓	✗	✓	✓	✓
2023.06	Orca [378]	Microsoft	Eng.	13B	N/A	Dec.-Only	NTP	BPE	RoPE	✓	✗	✗	✗	✓
2023.07	LLaMA 2 [557]	Meta	Eng.	70B	2T	Dec.-Only	NTP	BPE	RoPE	✓	✗	✓	✓	✓

Data Modalities for GenAI Models

Input: Text

Output: Text

- Sentence Completion
- Translation
- Question Answering
- Grammar Correction
- Conversations
- Summarization
- Creative Content Creation
- Code Generation

Output: Image

- Create Images
- Create Videos
- Create Animations

Output: Audio

- Create Audio from Text
- Audio Transcript in a Different language

Output: Decisions

- Solving Maths
- Game Playing

Data Modalities for GenAI Models

Input: Image

Output: Text

Visual Question Answering
Image Captioning
Image Search

Output: Image

Completing missing parts
of an image
Improving poor quality images
Enhancing resolution, colour etc.

Output: Video

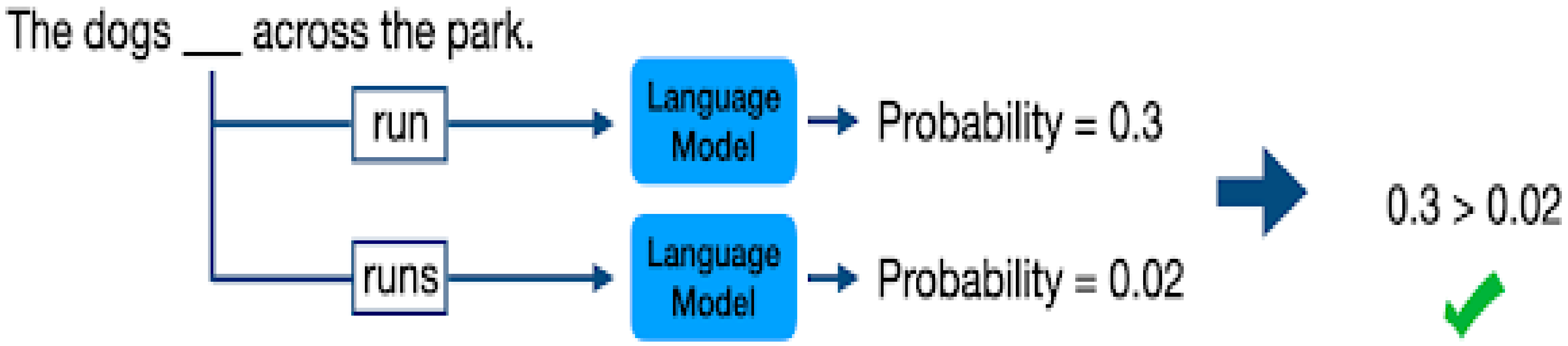
Creating videos from a set
Of images

How does generative AI work?

- Generative AI models use deep **neural networks** to identify the patterns and structures within existing data to generate new and original content.
- GenAI models can leverage different learning approaches – **Supervised, unsupervised or semi-supervised learning for training.**
- GenAI models often have a pre-trained **foundational model** that can be **finetuned** to a wide range of specific tasks down the line.
- E.g., popular applications like ChatGPT is based on the foundational model called GPT3.5. Stable Diffusion image model allows users to generate photorealistic images given a text input.

Language Model

- A Language model is a digital representation of a natural language's nuances.
- It a probability model that predicts what word comes next in a sequence of words. We train these models on large volumes of text, so they better understand what word is likely to come next.



Source: https://ai.googleblog.com/2021/12/evaluating-syntactic-abilities-of.html

A Language Model is a Probability Distribution over Strings of Text in a Language

- Probability distribution over strings of text
  - how likely is a given string (observation) in a given "language"
  - for example, consider probability for the following four strings
  - English: $p_1 > p_2 > p_3 > p_4$

P_1 = P("a quick brown dog")

P_2 = P("dog quick a brown")

P_3 = P("un chien quick brown")

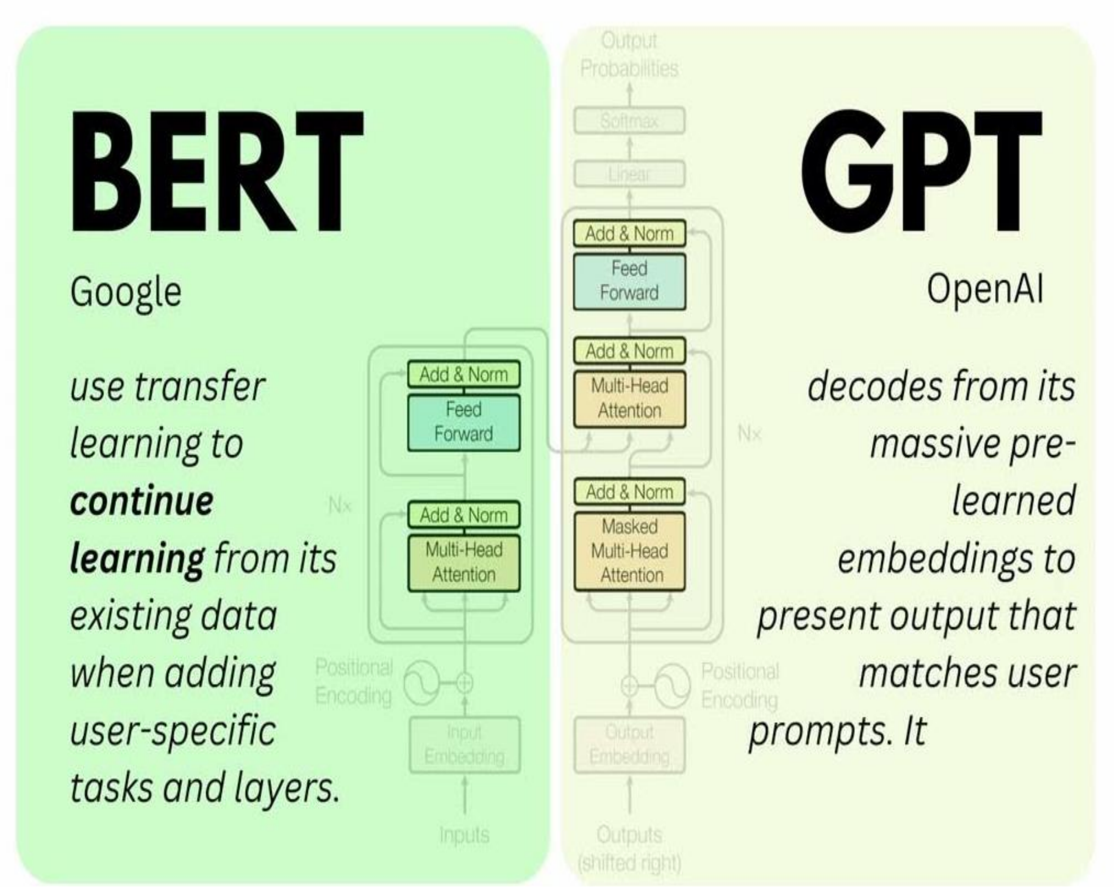
P_4 = P("un chien brun rapide")

  - … depends on what "language" we are modeling

Source: https://www.youtube.com/watch?v=W0TcVrI_vRg

Large Language Models

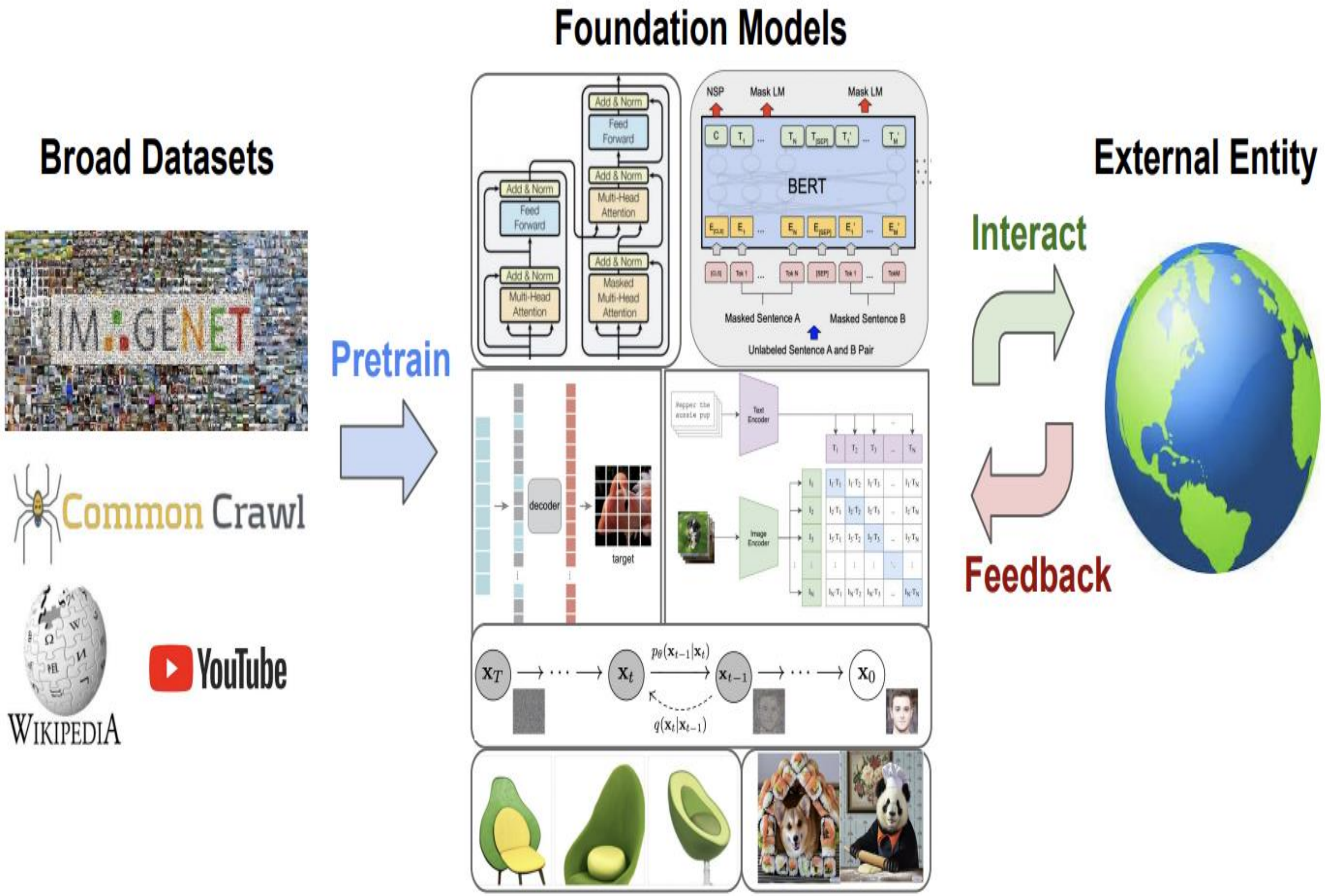
- Large language models(LLM) are generative models that work with language.
- They are called “large” as they have been trained on trillions of words over many weeks and months, and with large amounts of compute power.
- These foundation models with billions of parameters, can not only understand and generate natural language but can be used to break down complex tasks, reason, and solve problems.
- E.g., OpenAI’s GPT-4, Google’s PaLM, Meta’s LLaMA.



Source: https://www.youtube.com/watch?v=ewjlmLQI9kc

Foundational Models

- Foundational models are **neural networks** trained on massive **unlabelled datasets** that can be **fine-tuned** to perform a wide variety of tasks from translating text to analysing medical images.



Source: https://www.marktechpost.com/2023/03/11/the-scope-of-foundation-models-in-decision-making-their-challenges-opportunities-and-potentials/

Open Access and Closed Access Models

- Open access models are publicly available – Code, training data, steps, model weights, complete documentation. (GPT-Neo, GPT-J,OPT, FLAn-T5, StarCode, Falcon, INCITE, UL2, BLOOM)
- Closed access models may not reveal any of above. (Megatron TBLG, Galactica, Gopher, Chinchilla)
- Limited access models are intermediate. (GPT3, Jurrasic, CoHere, PaLM, ChatGPT, LLaMA, GPT-4)

https://stablediffusionweb.com

Making your dreams come true

Create Stable Diffusion images from text.

Easy to use

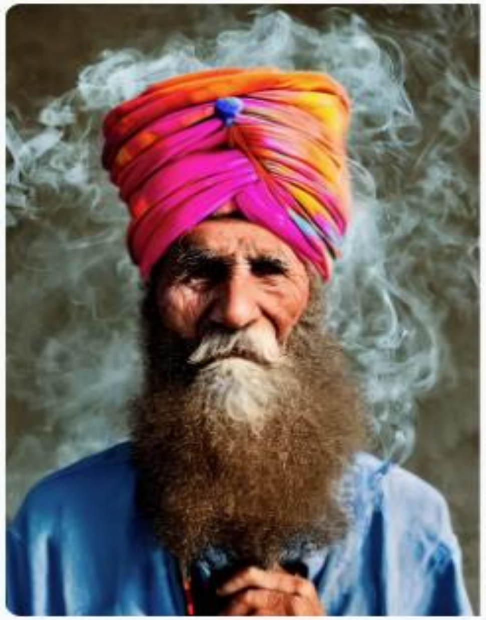
stablediffusionweb.com is an easy-to-use interface for creating images using the recently released Stable Diffusion image generation model.

High quality images

It can create high quality images of anything you can imagine in seconds–just type in a text prompt and hit Generate.

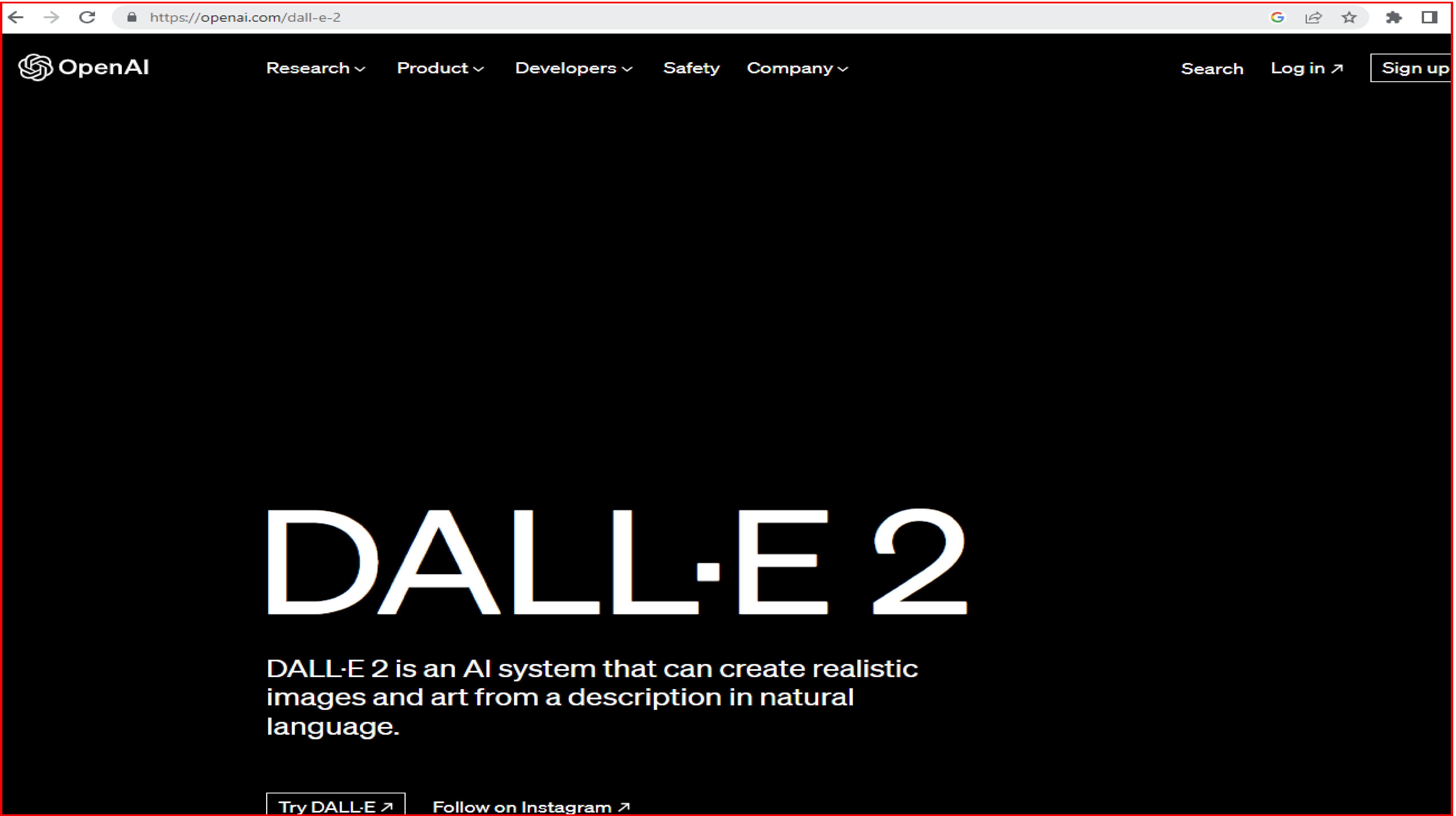
GPU enabled and fast generation

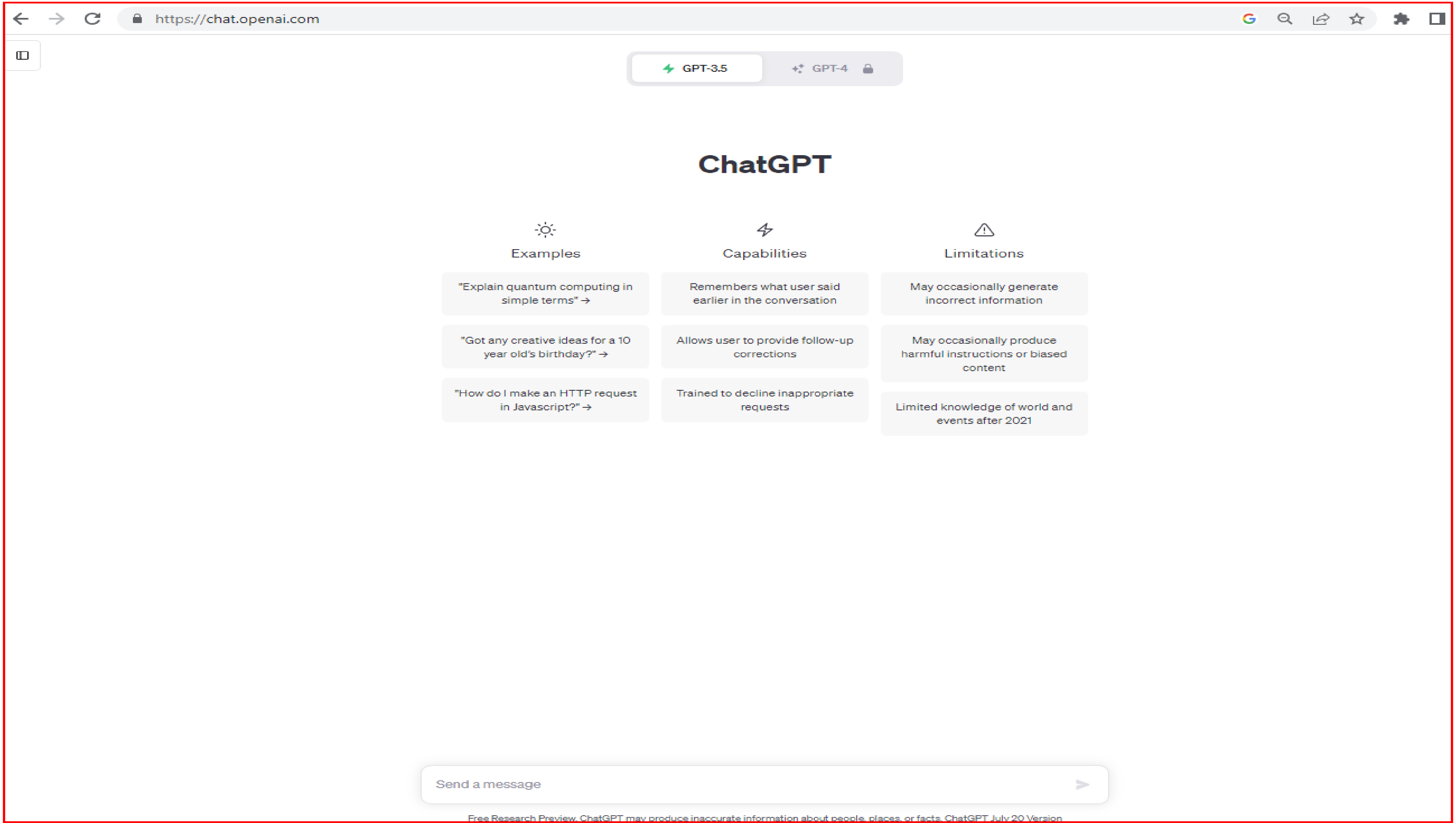
Perfect for running a quick sentence through the model and get results back rapidly.

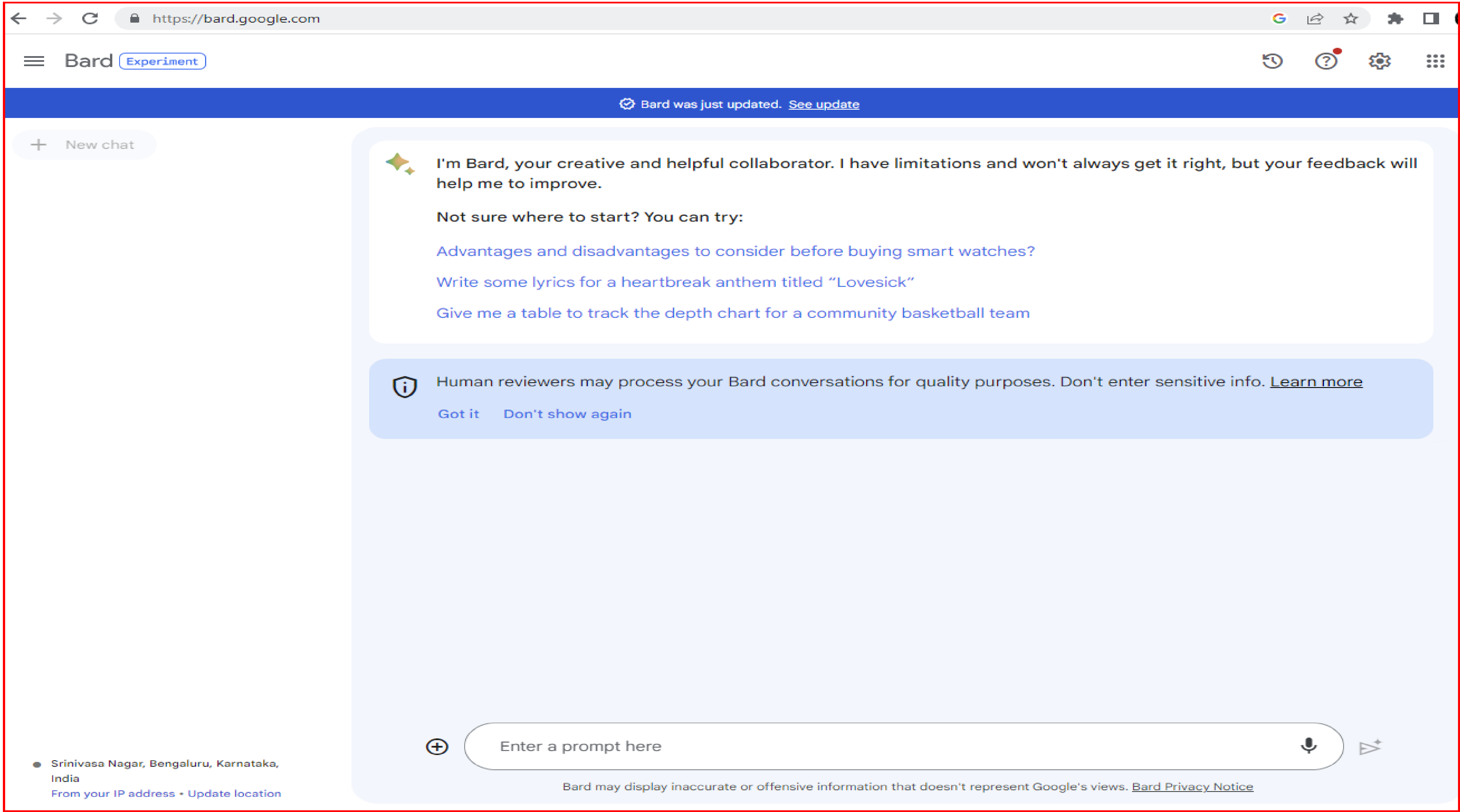


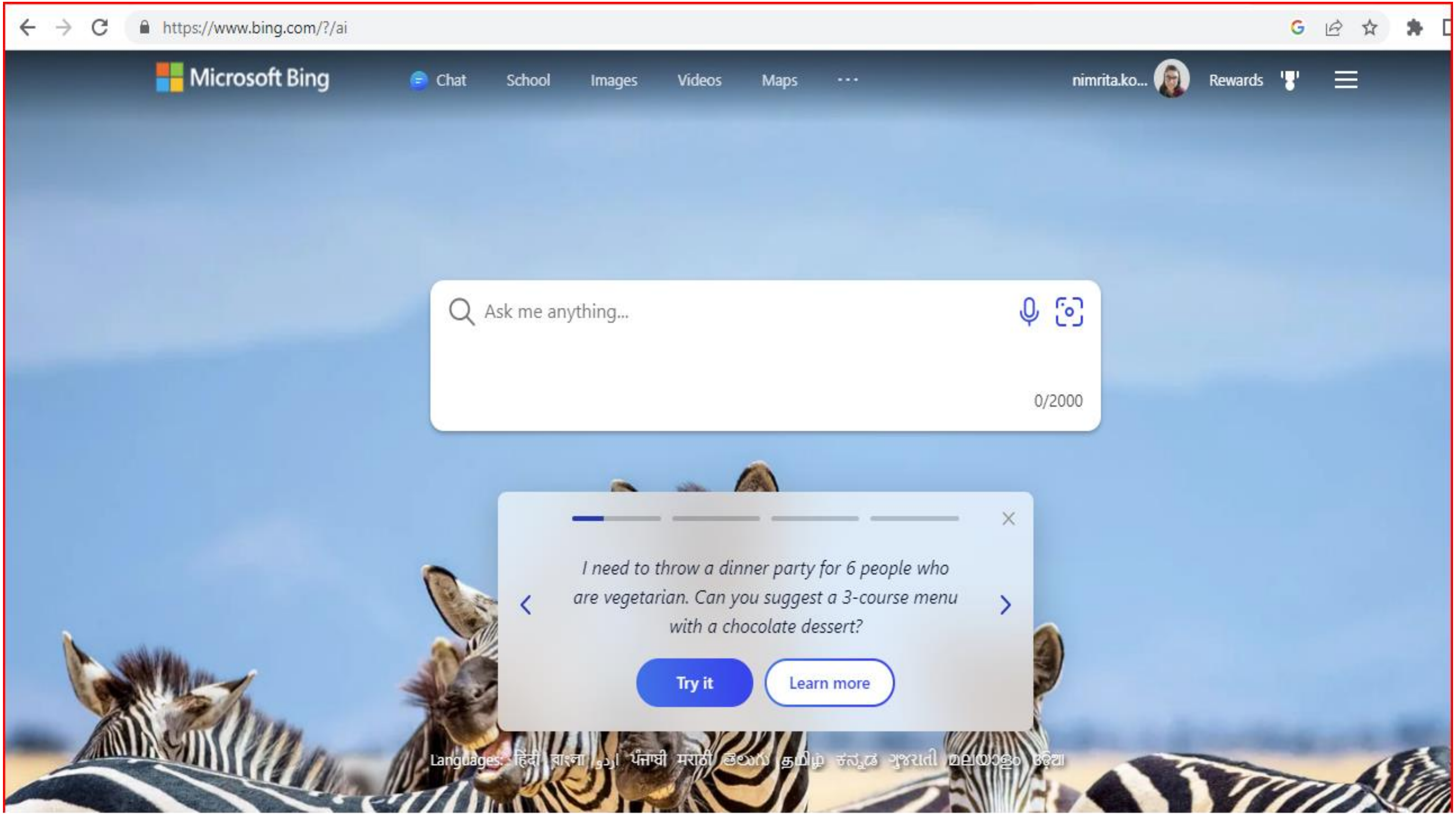
Privacy

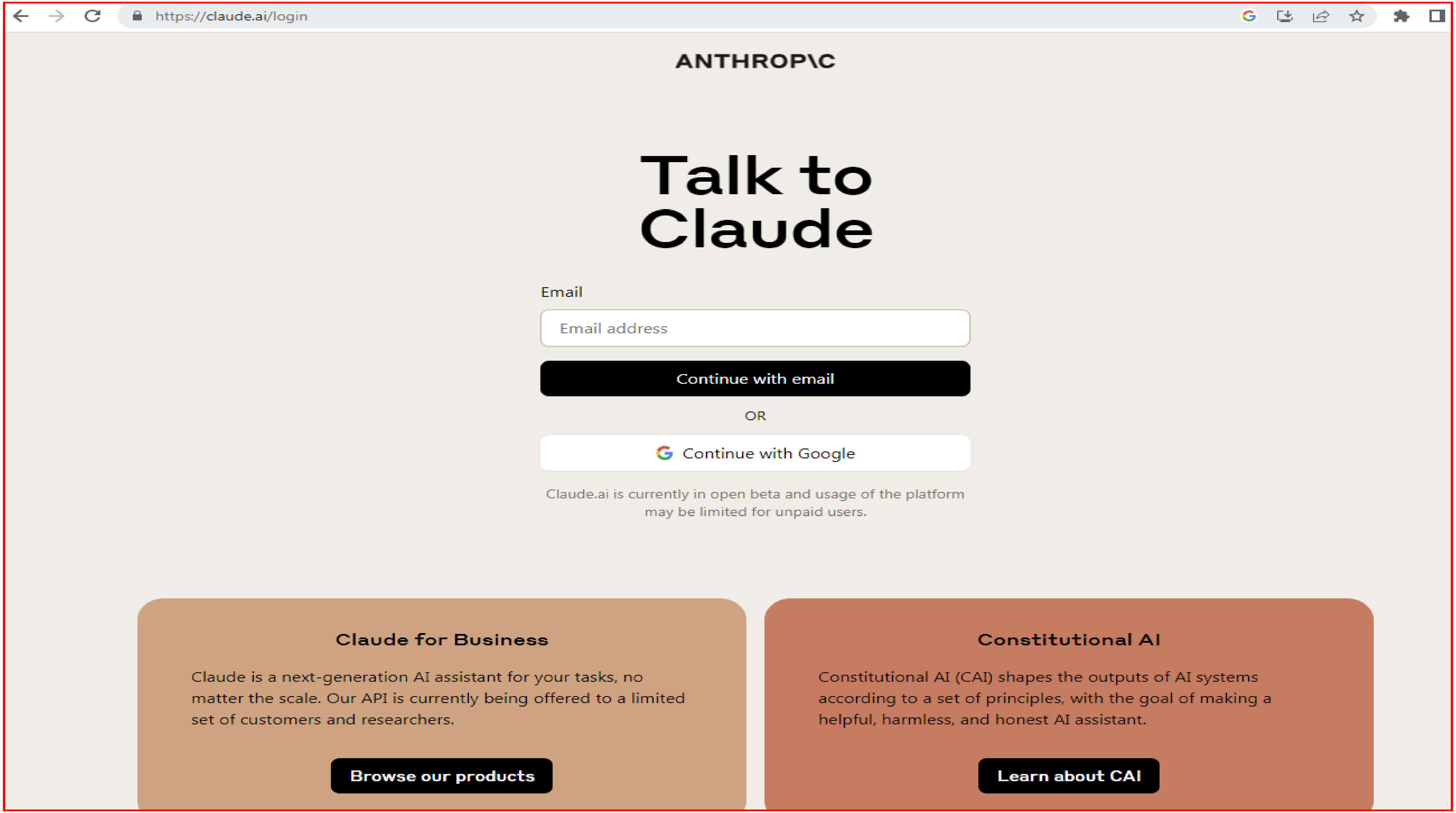
We case about your privacy.


https://openai.com/dall-e-2
OpenAI
Research
Product
Developers
Safety
Company
Search
Log in ↗
Sign up
DALL·E 2
DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.
Try DALL·E ↗
Follow on Instagram ↗


https://chat.openai.com
GPT-3.5
GPT-4
ChatGPT
Examples
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
Limitations
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021
Send a message
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT July 20 Version


https://bard.google.com
Bard Experiment
Bard was just updated. See update
New chat
I'm Bard, your creative and helpful collaborator. I have limitations and won't always get it right, but your feedback will help me to improve.
Not sure where to start? You can try:
Advantages and disadvantages to consider before buying smart watches?
Write some lyrics for a heartbreak anthem titled "Lovesick"
Give me a table to track the depth chart for a community basketball team
Human reviewers may process your Bard conversations for quality purposes. Don't enter sensitive info. Learn more
Got it
Don't show again
Enter a prompt here
Srinivasa Nagar, Bengaluru, Karnataka, India
From your IP address • Update location
Bard may display inaccurate or offensive information that doesn't represent Google's views. Bard Privacy Notice


https://www.bing.com/?/ai
Microsoft Bing
Chat
School
Images
Videos
Maps
nimrita.ko...
Rewards
Ask me anything...
0/2000
I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?
Try it
Learn more
Languages: हिंदी বাংলা اردو ਪੰਜਾਬੀ मराठी తెలుగు தமிழ் ಕನ್ನಡ ગુજરાતી മലയാളം ଓଡ଼ିଆ


https://claude.ai/login
ANTHROP\C
Talk to Claude
Email
Email address
Continue with email
OR
Continue with Google
Claude.ai is currently in open beta and usage of the platform may be limited for unpaid users.
Claude for Business
Claude is a next-generation AI assistant for your tasks, no matter the scale. Our API is currently being offered to a limited set of customers and researchers.
Browse our products
Constitutional AI
Constitutional AI (CAI) shapes the outputs of AI systems according to a set of principles, with the goal of making a helpful, harmless, and honest AI assistant.
Learn about CAI

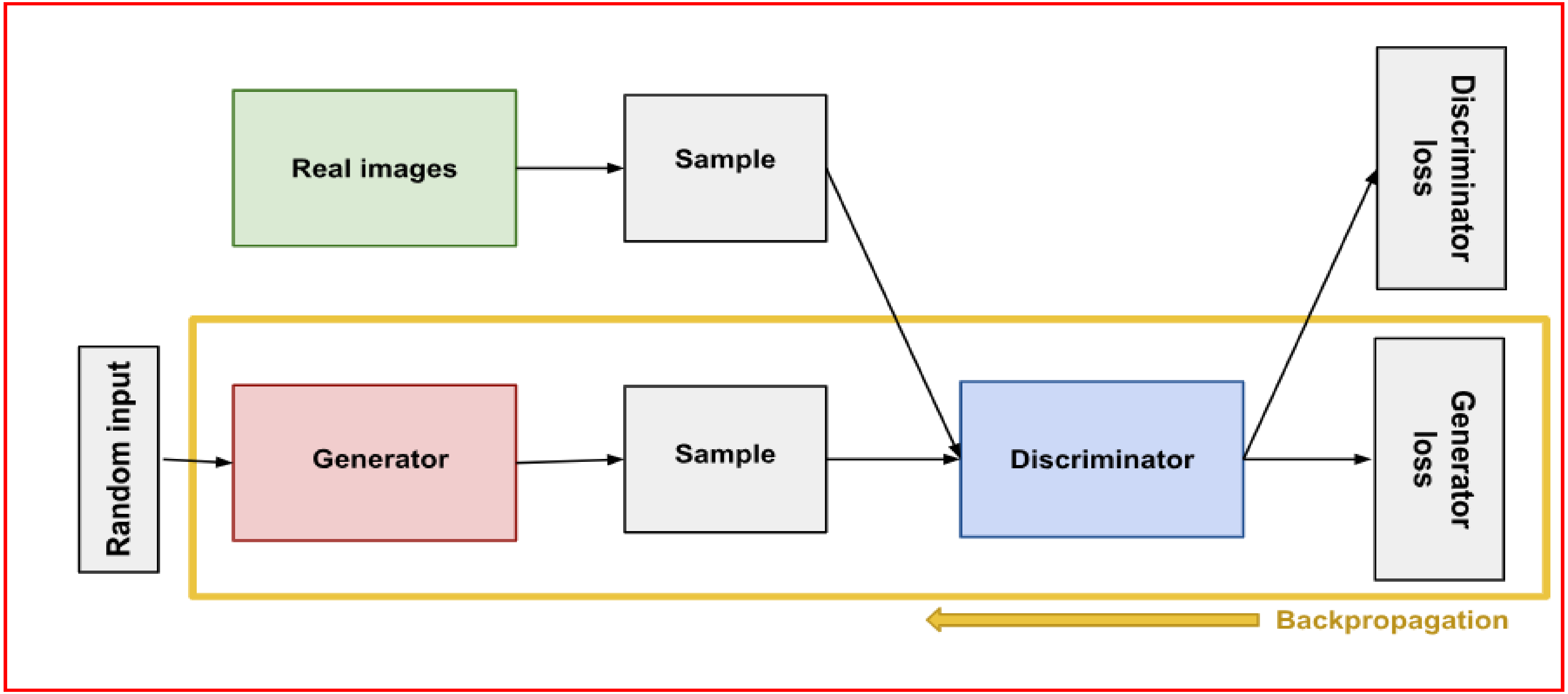
Deep Neural Architectures used by Gen AI Models

- The GenAI models combine various AI algorithms.
- First, the input prompt is represented in a numerical vector (Embeddings).
- Then a neural network is used to generate new content in response to the prompt. The neural architecture used depends on the content type that the model needs to generate.
- Deep Neural Network Architectures like **GANs and Variational Autoencoders** (VAEs) are used for generating realistic human faces, synthetic data for AI training. **Transformers** are used for text, images generation.

Generative Adversarial Network (GAN)

- A Generative Adversarial Network (GAN) is a generative modeling method used for text to image generation, image super resolution, denoising.
- A GAN contains two sub-models that compete and feed off each other to produce more realistic outputs:
  - The generator model—trained to generate new outputs. The generator attempts to fool the discriminator and trains on more data to produce plausible results.
  - The discriminator model—classifies inputs as realistic or fake.
- This adversarial approach helps to improve the generator model’s capabilities until the discriminator model cannot distinguish between real and generated inputs. Through backpropagation, the discriminator's classification provides a signal that the generator uses to update its weights.

GAN Architecture



Source: https://developers.google.com/machine-learning/gan/generator

Variational Autoencoders

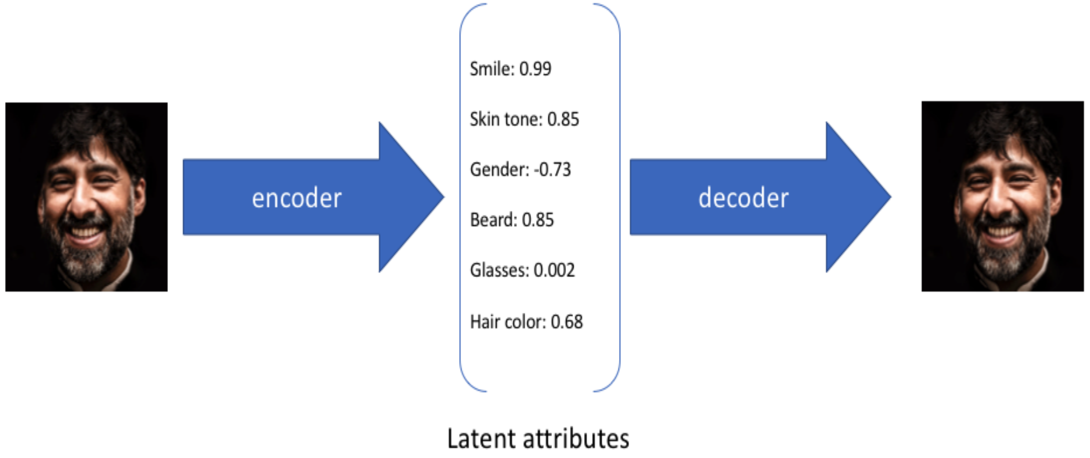
Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

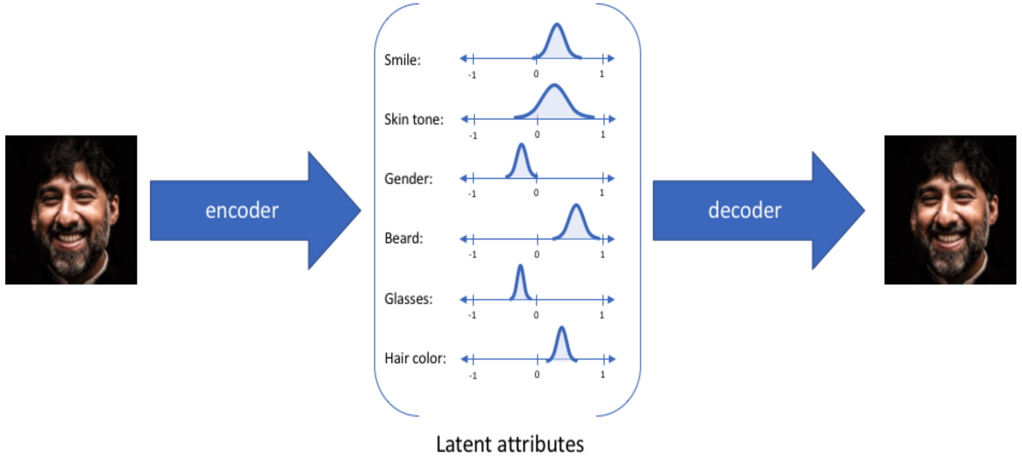
Autoencoders

- An autoencoder network is a pair of connected networks, an encoder and a decoder.
- An encoder network takes in an input, and converts it into a smaller, dense representation, called as encoding vectors (latent representation), which the decoder network can use to convert it back to the original input.
- Each dimension in the encoding vector represents some learned attribute of the data.
- The decoder takes these vectors and reconstructs the input data from them.
- The latent space of encoded vectors may not be continuous.

Variational Autoencoder (VAE)

- Unlike autoencoders, the latent spaces of variational autoencoders are continuous, allowing easy random sampling and interpolation.
- In VAE, the encoder outputs two vectors of size n, a vector of means μ and a vector of standard deviations σ.

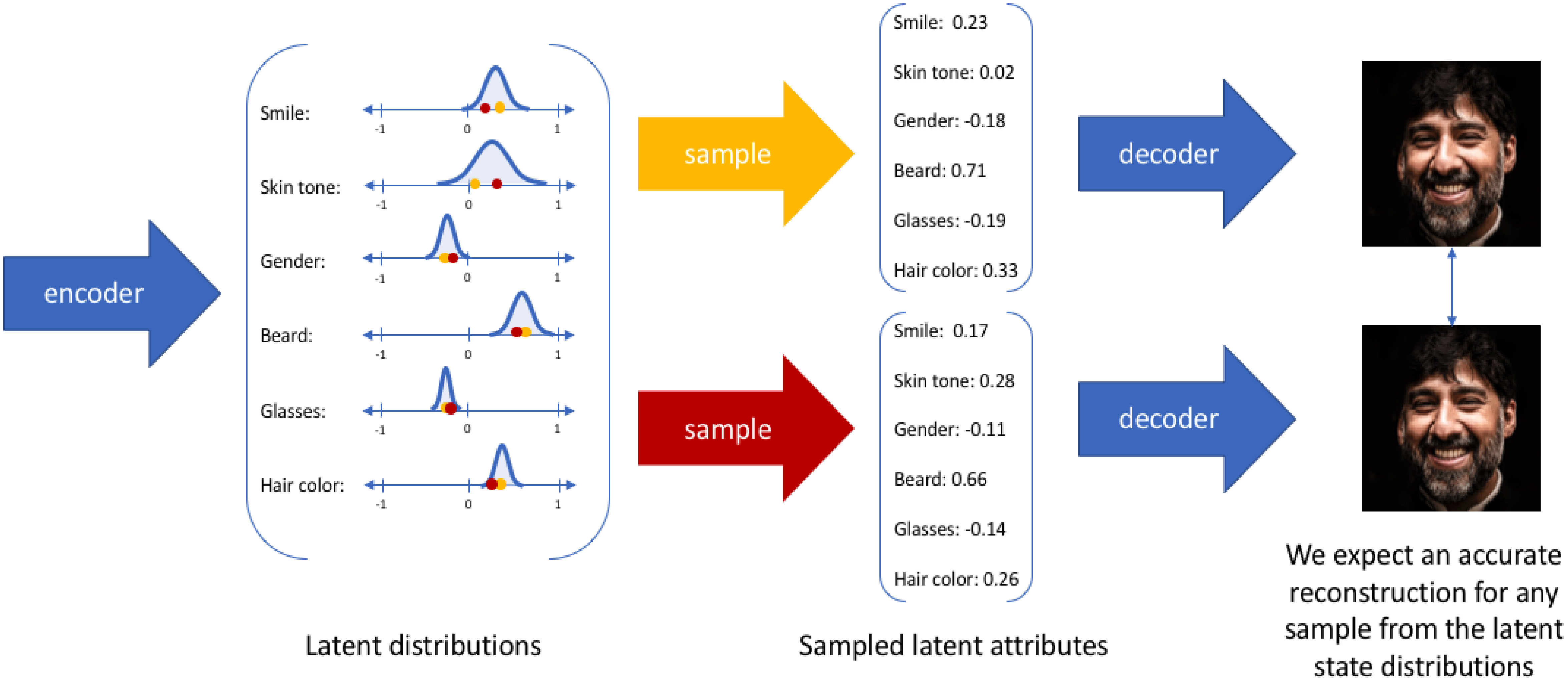




Latent representations in autoencoders are non-continuous

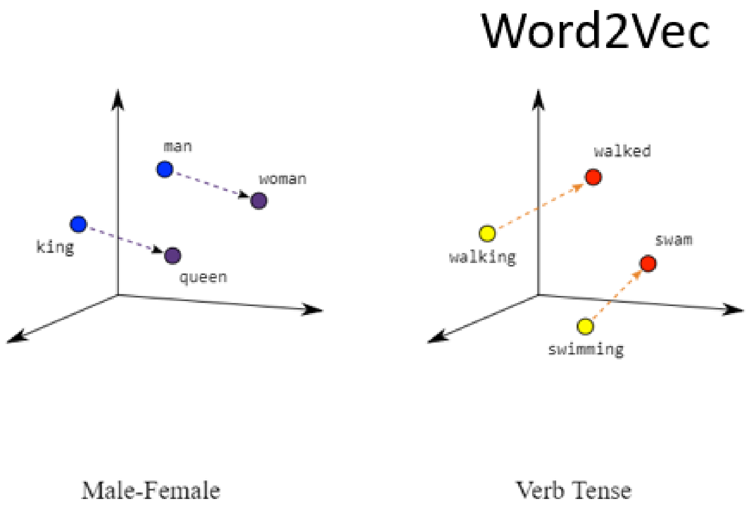
Latent representations in VAE are continuous

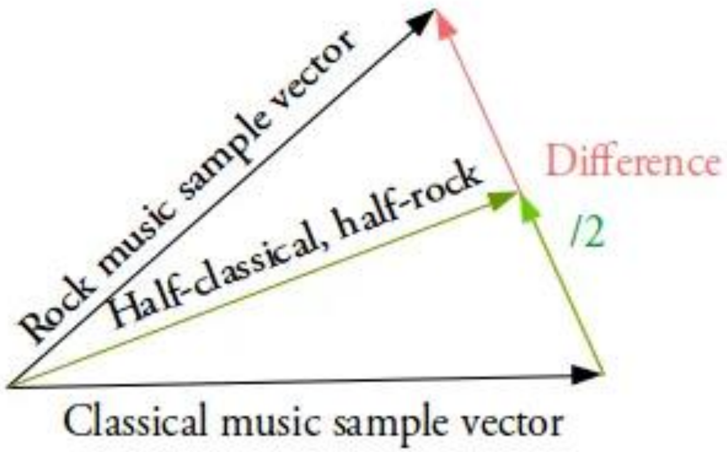
Source: https://www.jeremyjordan.me/variational-autoencoders/

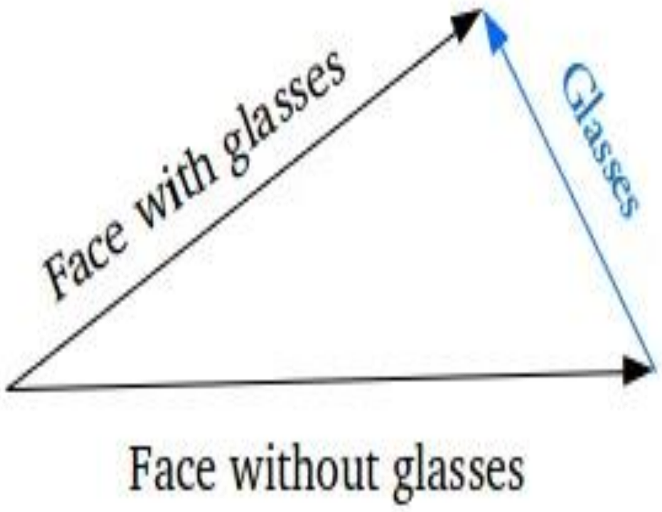


Source: https://www.jeremyjordan.me/variational-autoencoders/

Word Embeddings and their relationships

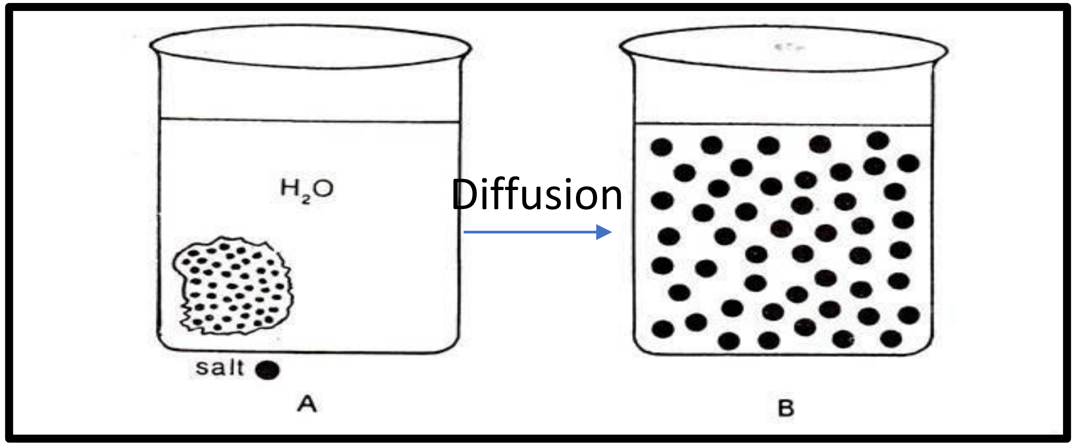


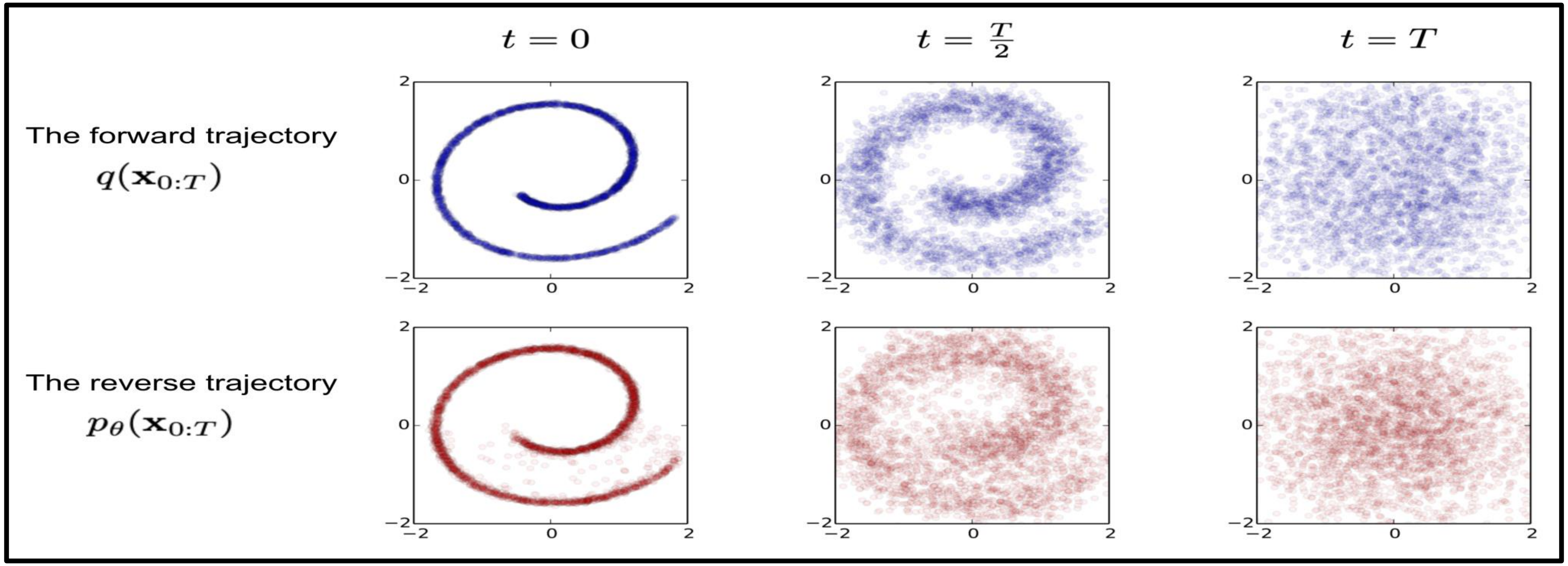




https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Image Generation by Reverse Diffusion



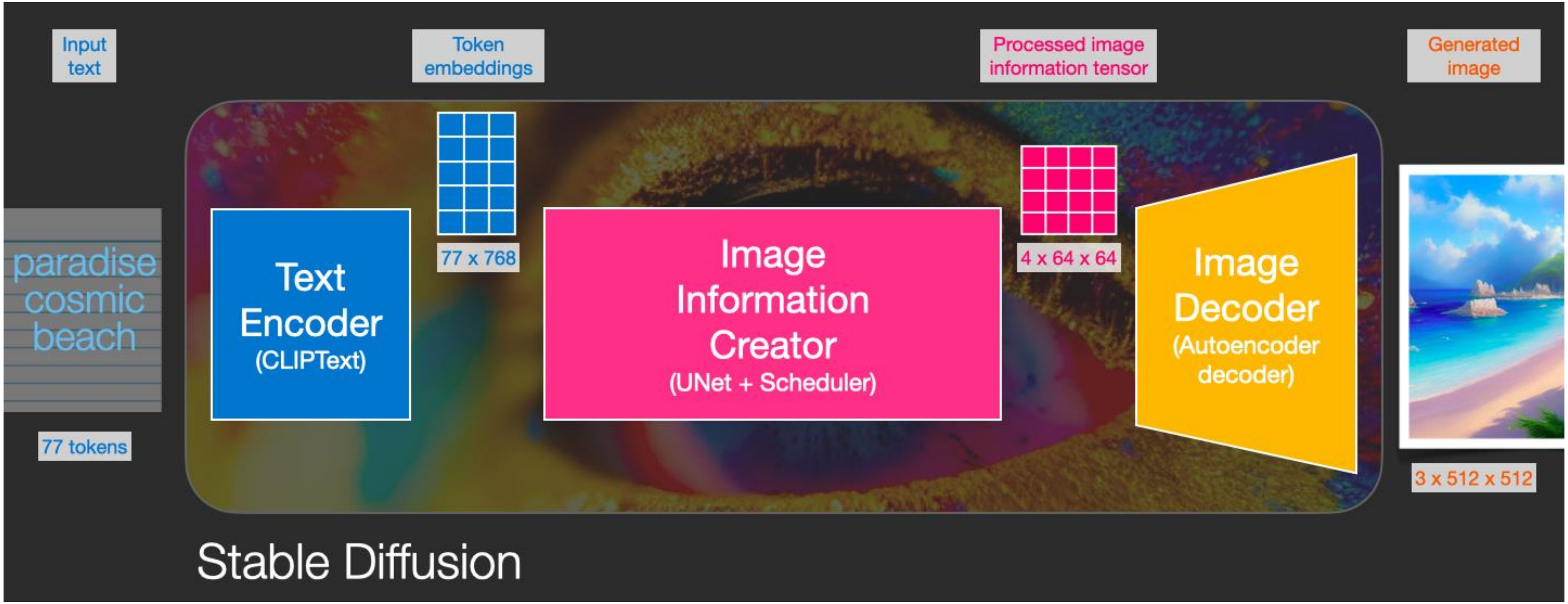


Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Stable Diffusion

- Stable Diffusion is a system made up of **three main components** each with its own neural network.
- First, there is a **text-encoder (a Transformer model)** that translates the text information into a numeric vector (one vector per token) that captures the ideas in the text. That information is then presented to the Image Generator.
- Second is the **Image Generator.** It has two stages or sub parts– **image information creator and image decoder.**
- **Image information creator is a UNet neural network and a scheduling algorithm**. This component implements “Diffusion” i.e., gradually diffuses information in the latent space. It takes as input the text embeddings and a starting multidimensional array of noise and outputs processed array.
- **Image decoder is an autoencoder** that paints a final picture from the information it got from the information creator. It runs only once at the end of the process.

Stable Diffusion

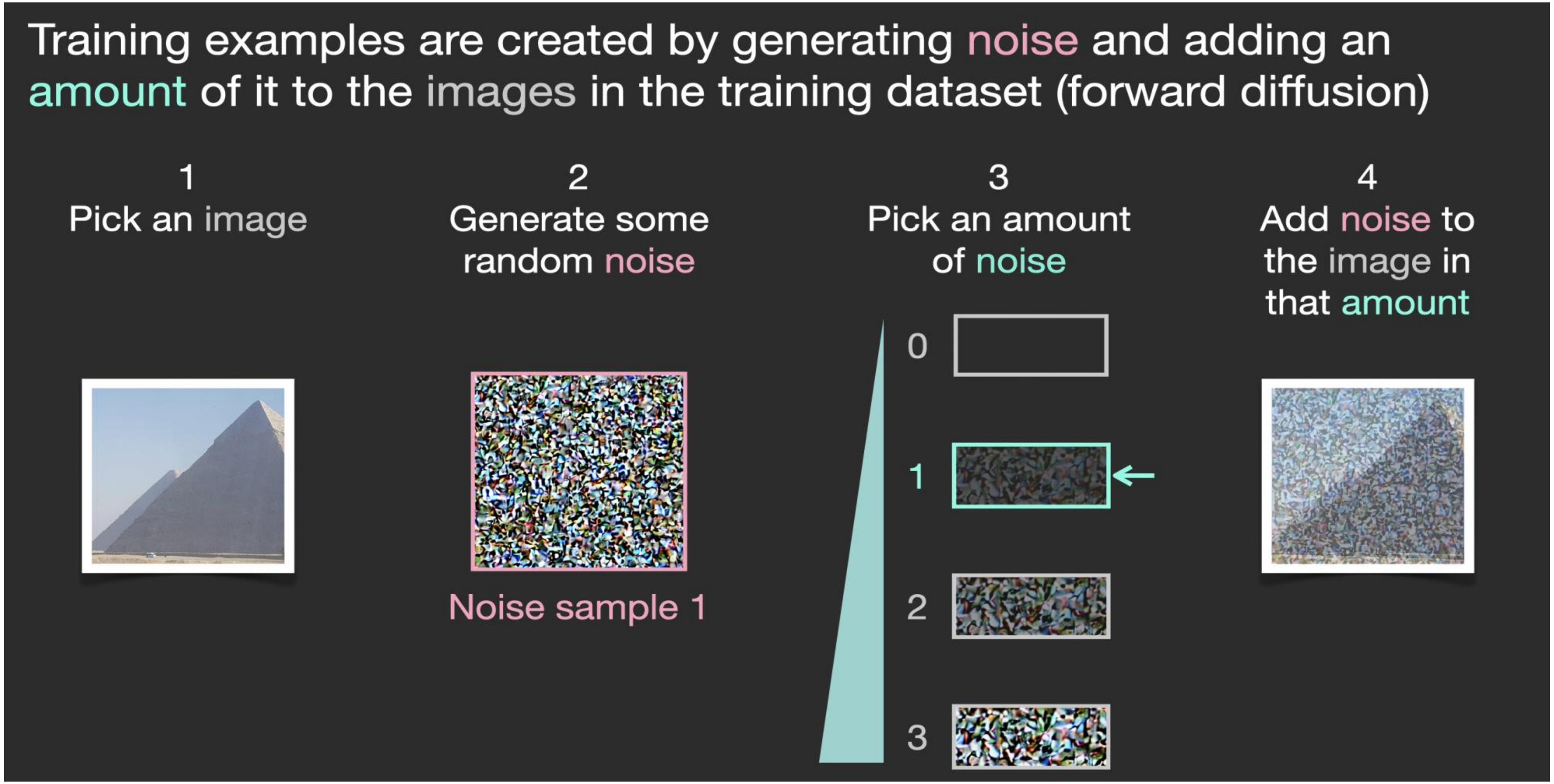


https://jalammar.github.io/illustrated-stable-diffusion/

Nimrita Koul

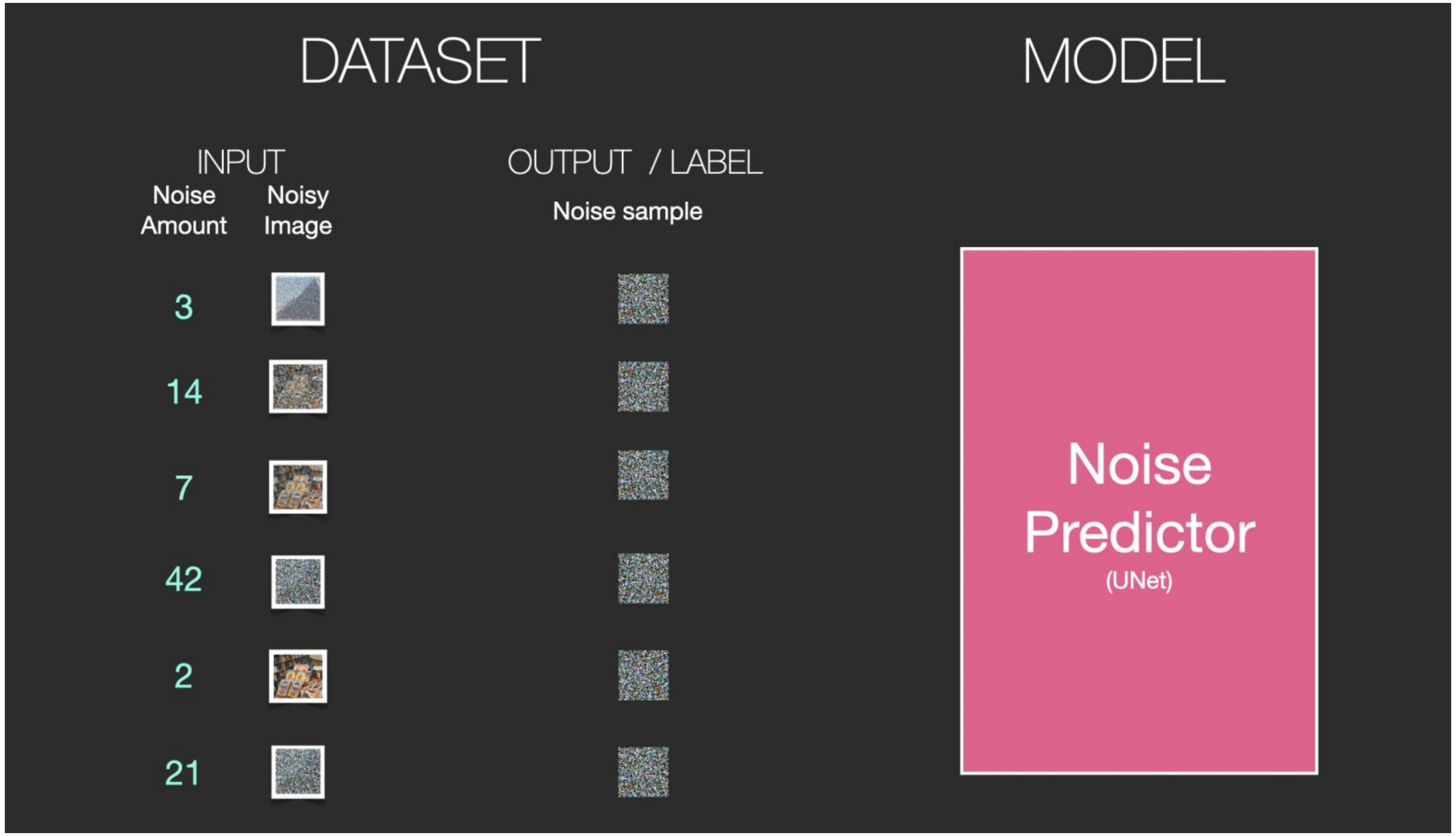
So, What is Diffusion?

- Say we have an image, we generate some noise, and add it to the image.



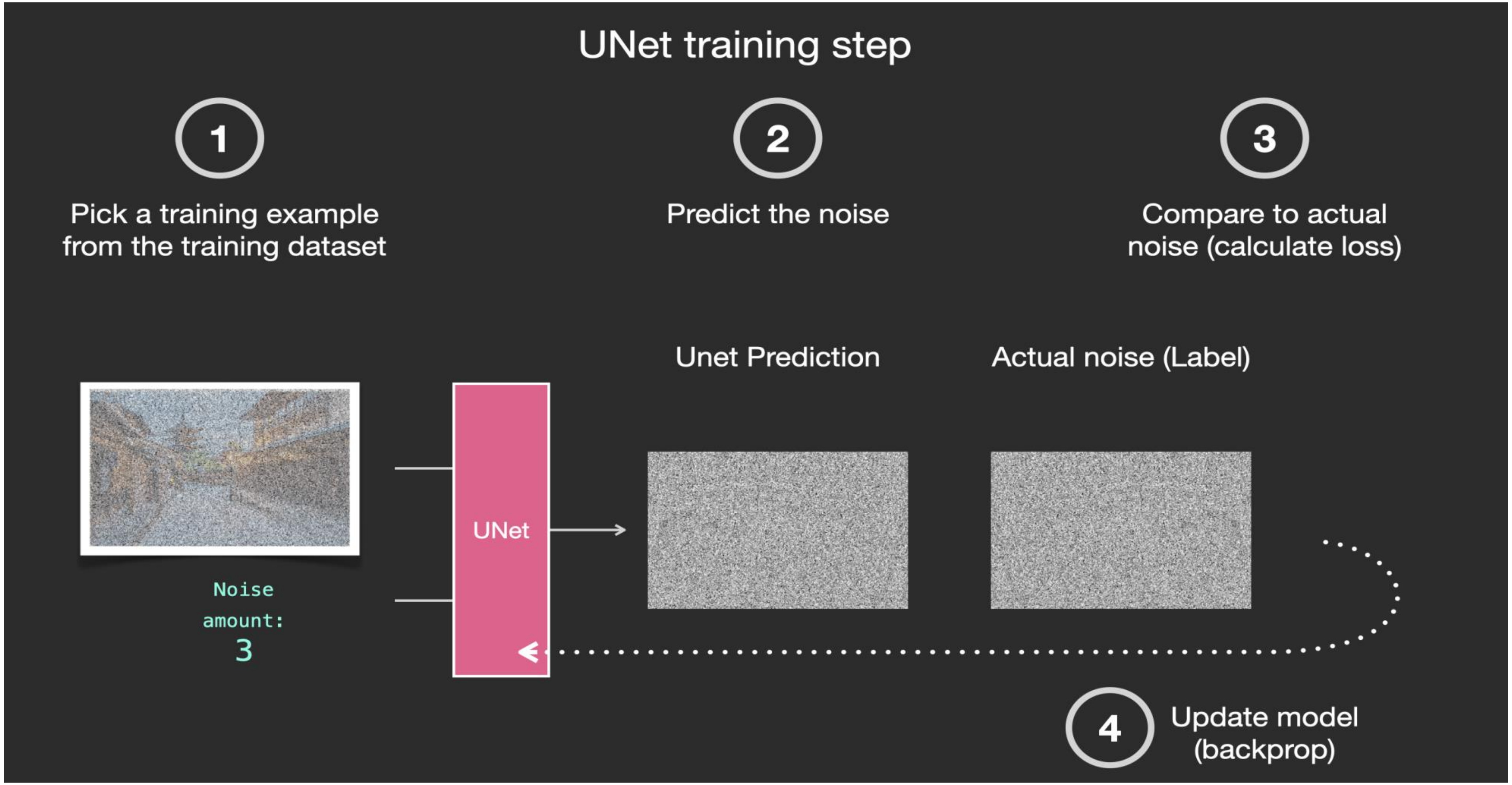
https://jalammar.github.io/illustrated-stable-diffusion/

- This image with added noise is now a new training example. Similarly, we can create lots of training examples to train the image generation model and the noise predictor that creates images when run in a certain configuration



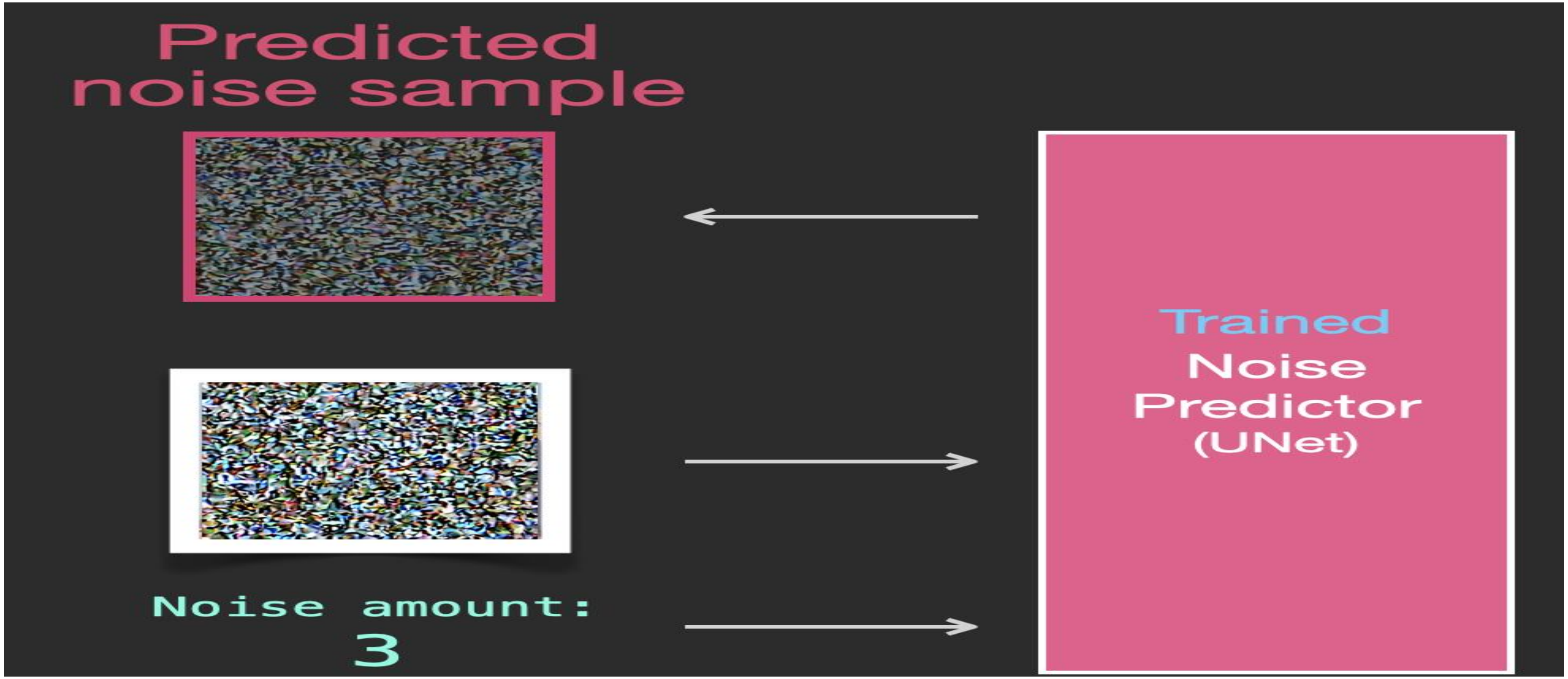
https://jalammar.github.io/illustrated-stable-diffusion/

- With this dataset, we can train the noise predictor that creates images when run in a certain configuration.



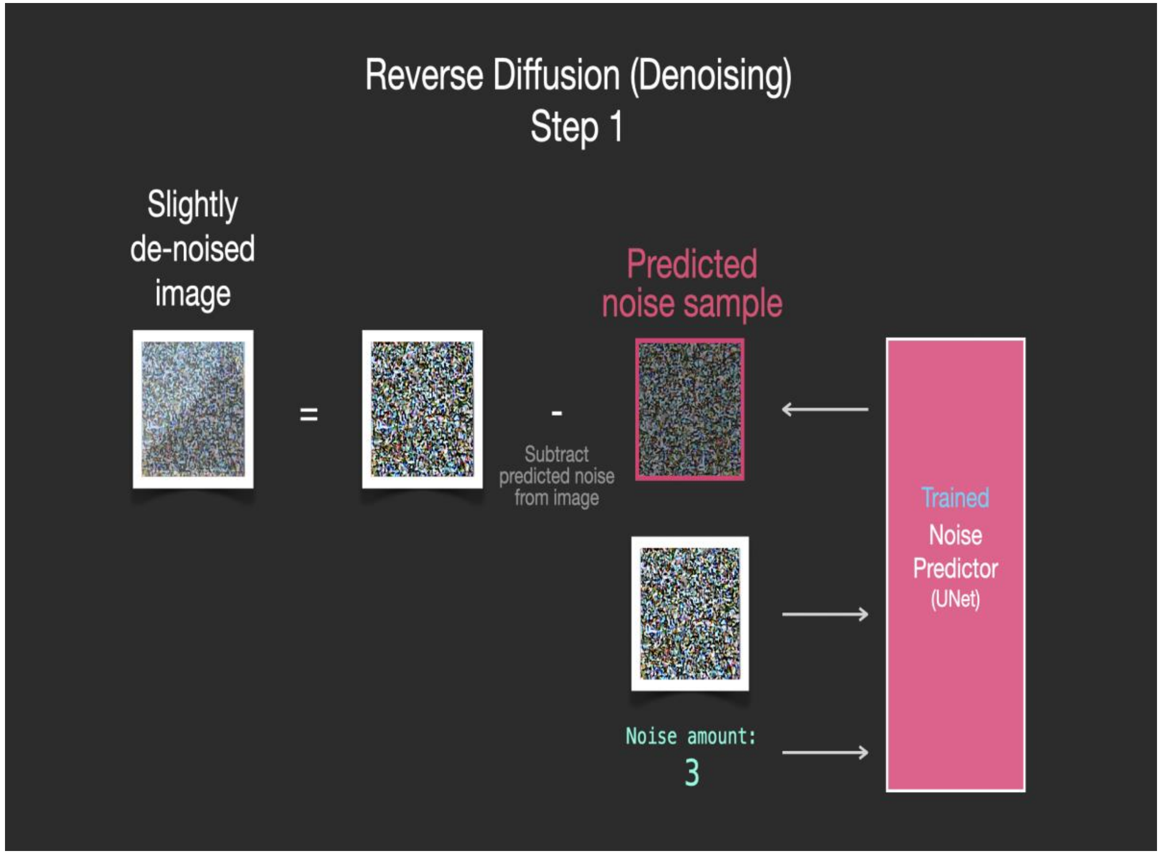
https://jalammar.github.io/illustrated-stable-diffusion/

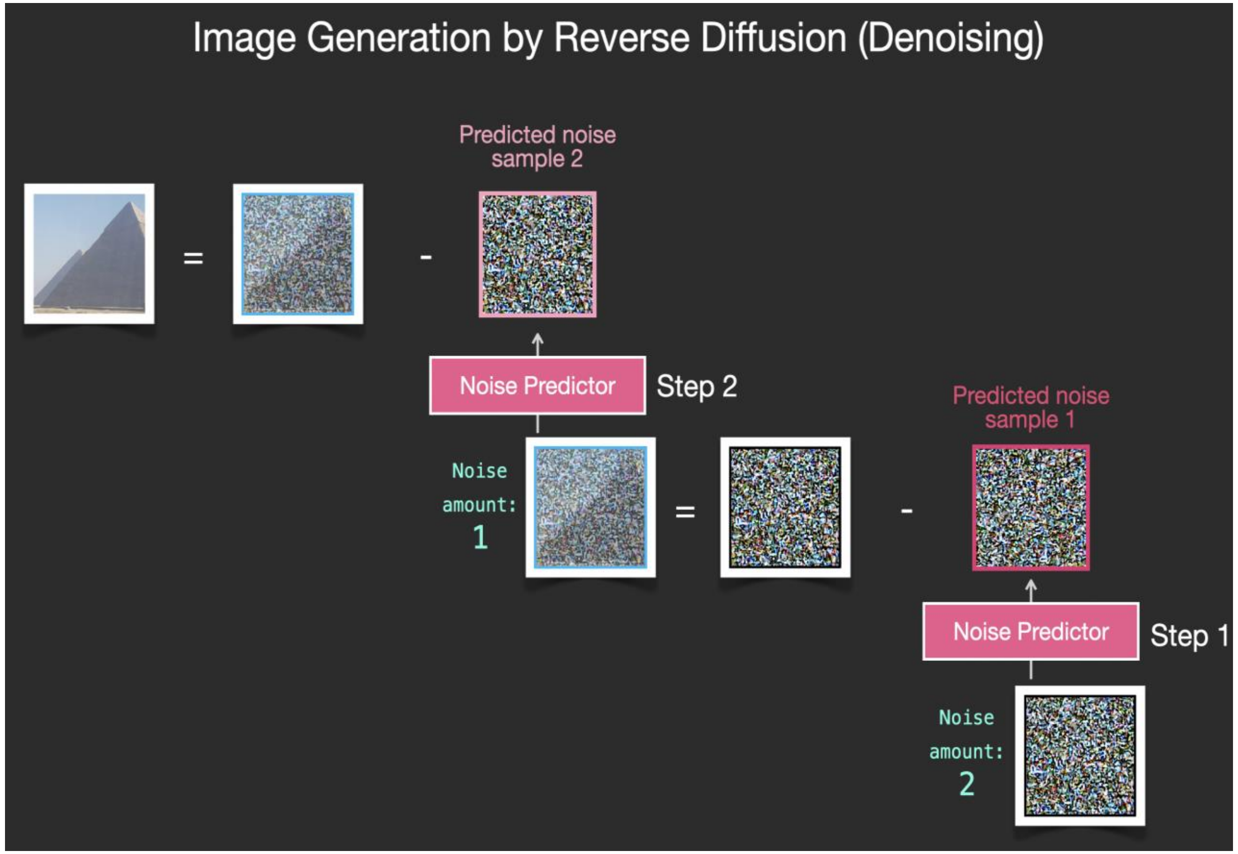
- The trained noise predictor can take a noisy image, and the number of the denoising step, and is able to predict a slice of noise.
- The predicted sampled noise can be subtracted from a given image to create an image that's closer to the images the model was trained on.

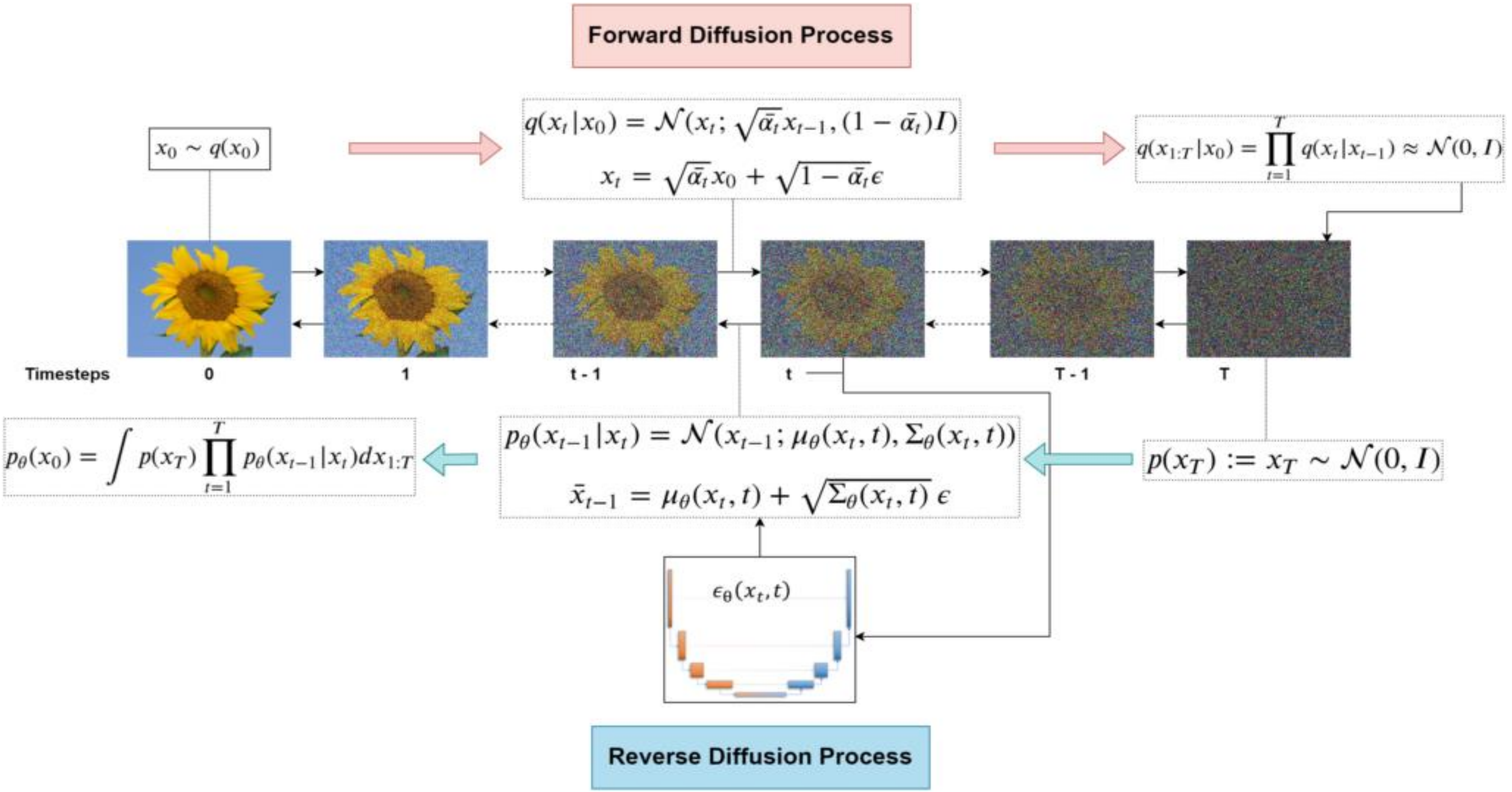


https://jalammar.github.io/illustrated-stable-diffusion/

Image Generation by Reverse Diffusion / Denoising

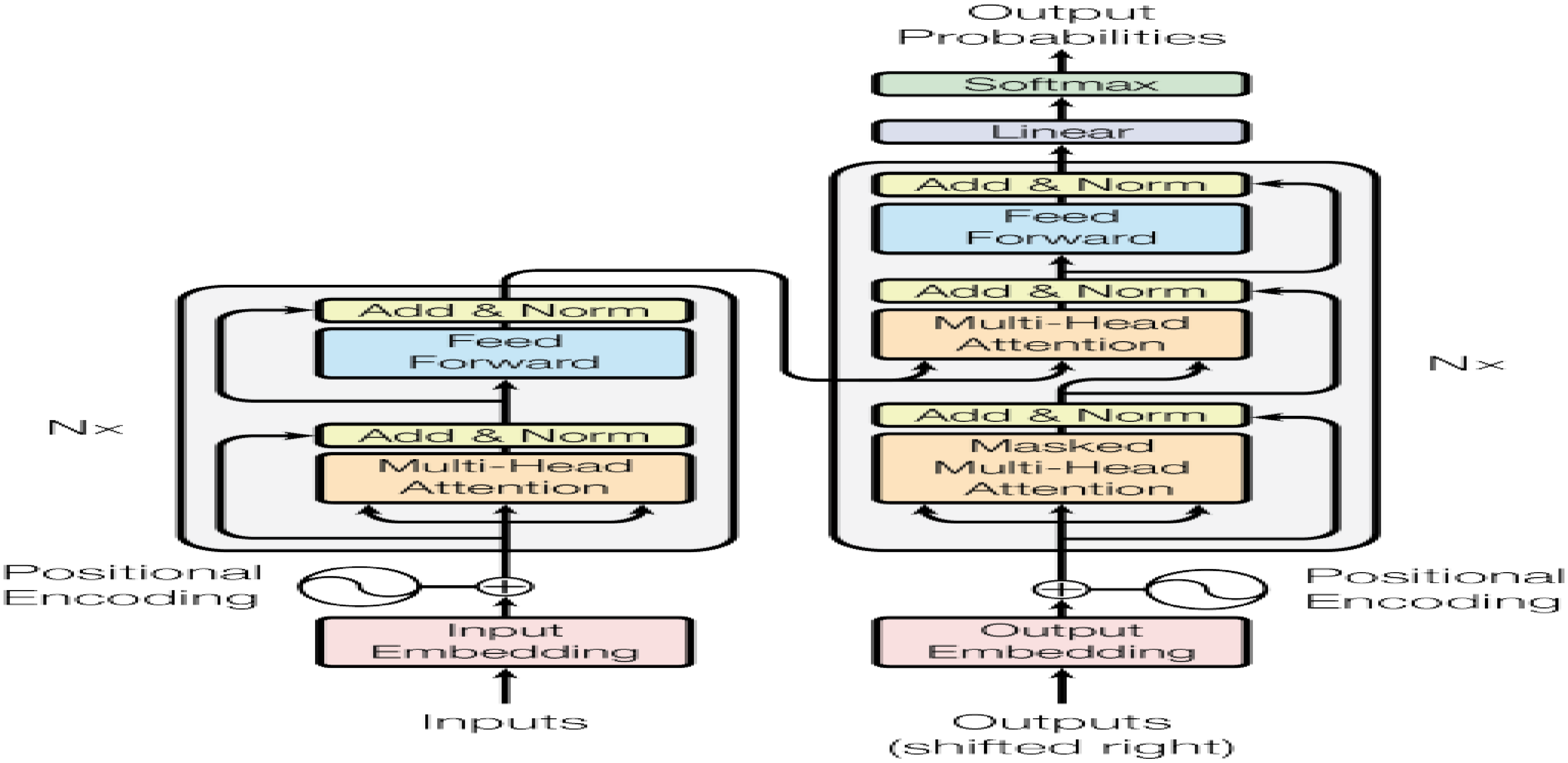






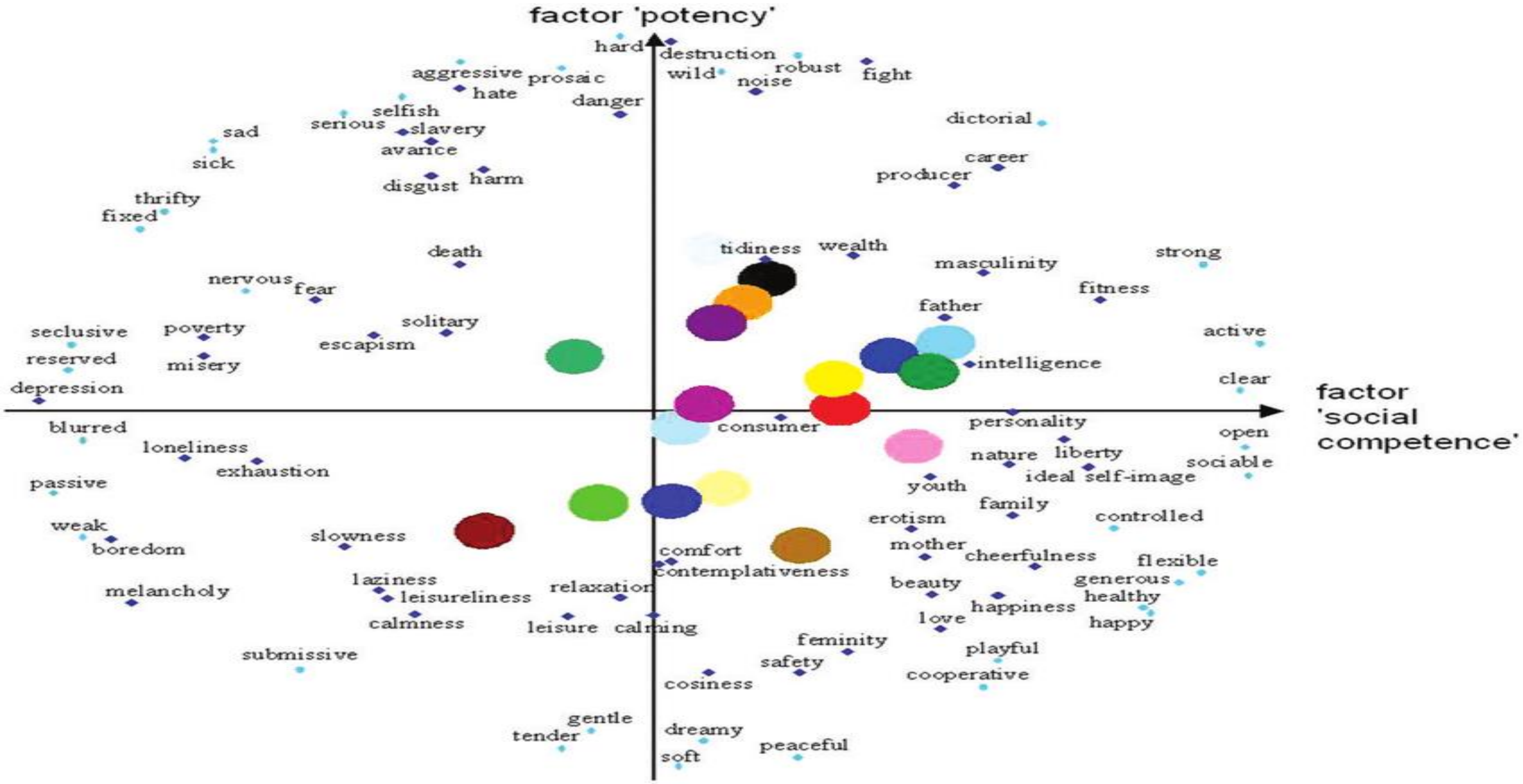
Source: https://learnopencv.com/denoising-diffusion-probabilistic-models/

Transformers: The Architecture Behind LLMs



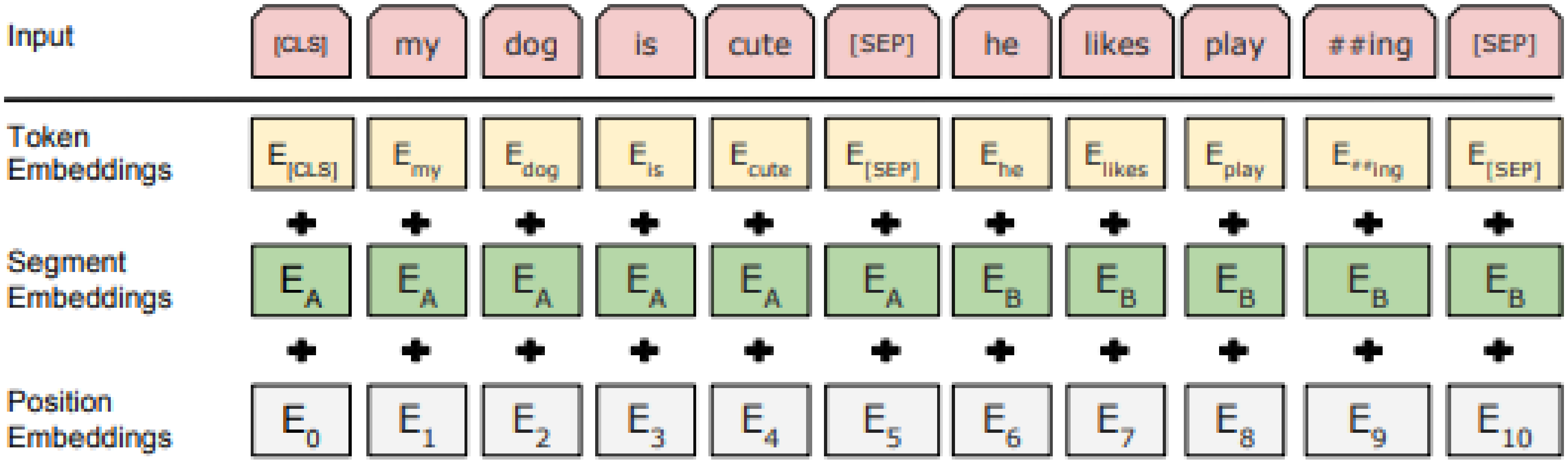
Source: Paper Attention is all you need. https://arxiv.org/abs/1706.03762

Word Embeddings



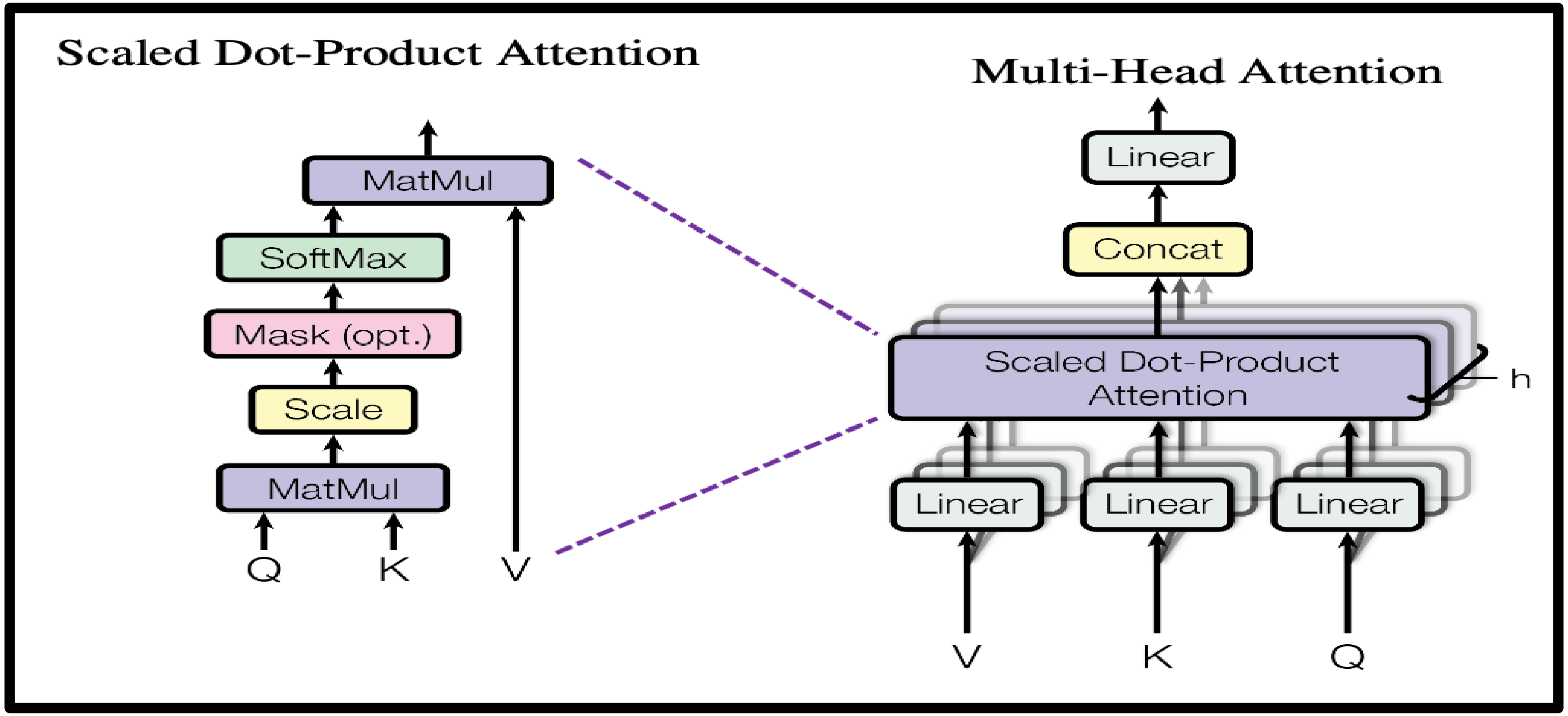
https://discuss.huggingface.co/t/get-word-embeddings-from-transformer-model/6929

Text Representation in Google's BERT

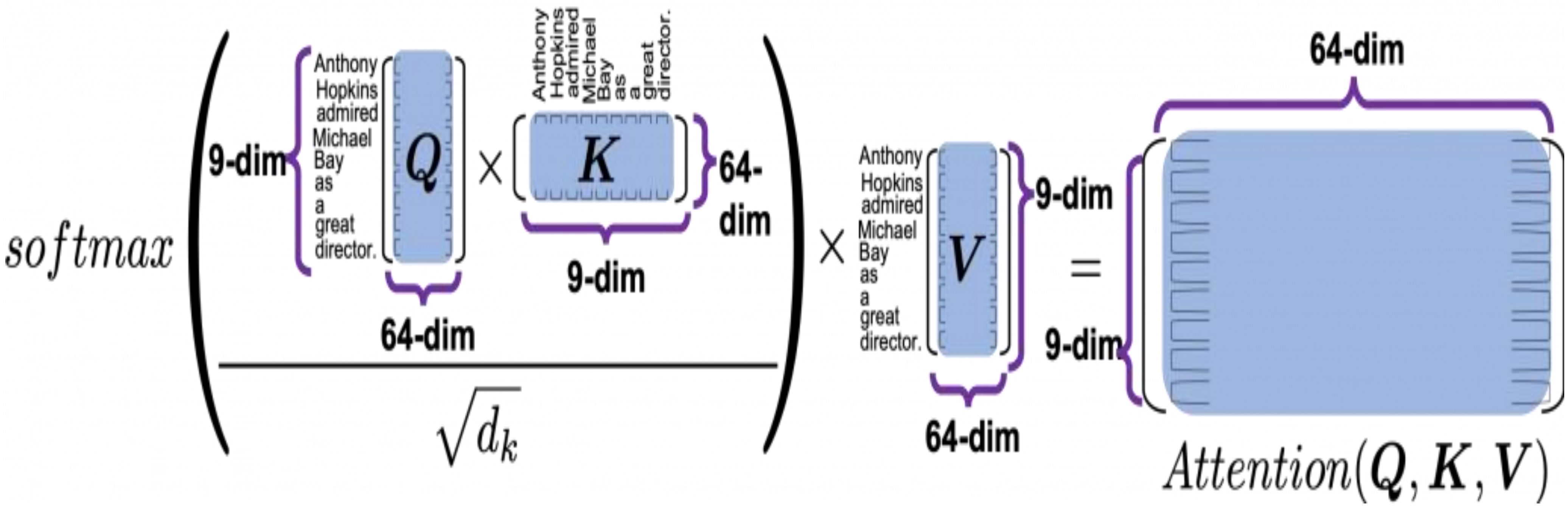


Text Embeddings in BERT:
Source: https://medium.com/@_init_/why-bert-has-3-embedding-layers-and-their-implementation-details-9c261108e28a

Attention Mechanism



Multiheaded Attention:
Source: https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/



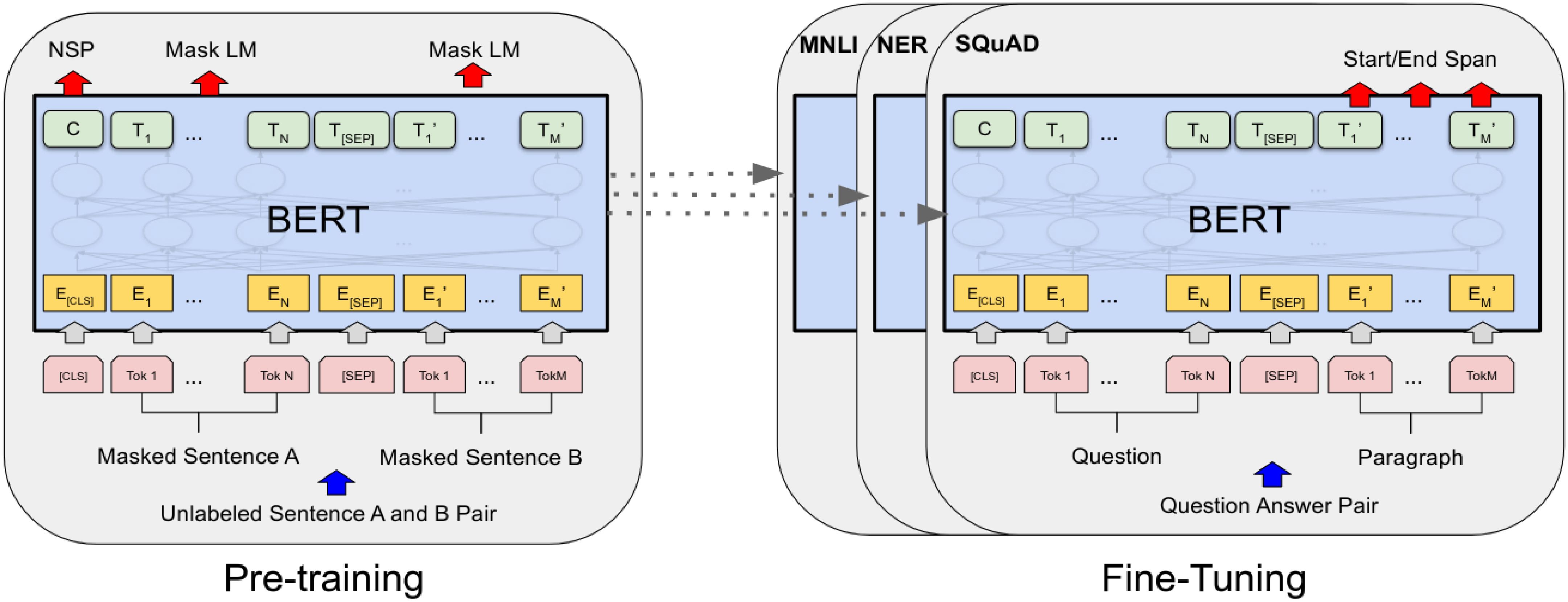
Source: https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/

We you compare the 'query' with the 'keys' and get scores/weights for the 'values.' Each score/weight is the relevance between the 'query' and each 'key'. And you reweight the 'values' with the scores and take the summation of the reweighted values.

Tasks that benefit from attention mechanism

1. Language Modelling
2. Semantic Segmentation
3. Information Retrieval
4. Object Detection
5. Sentiment Analysis
6. Text Generation
7. Question Answering
8. Text Classification
9. Image Classification

Google BERT's Pre-training and Fine Tuning



Source: https://paperswithcode.com/method/bert

Training Large Language Models

- Pre-Training
- In context Learning (Prompt Based Learning)
- Supervised Fine-Tuning (SFT)
- Reinforcement Learning from Human Feedback (RLFH)

Training LLMs: Pretraining

- During pretraining the model learns the nuances of the language and learns to predict the next word in a sequence of words.

Training LLMs: In-context Learning

During prompting the model learns to improve its outputs.

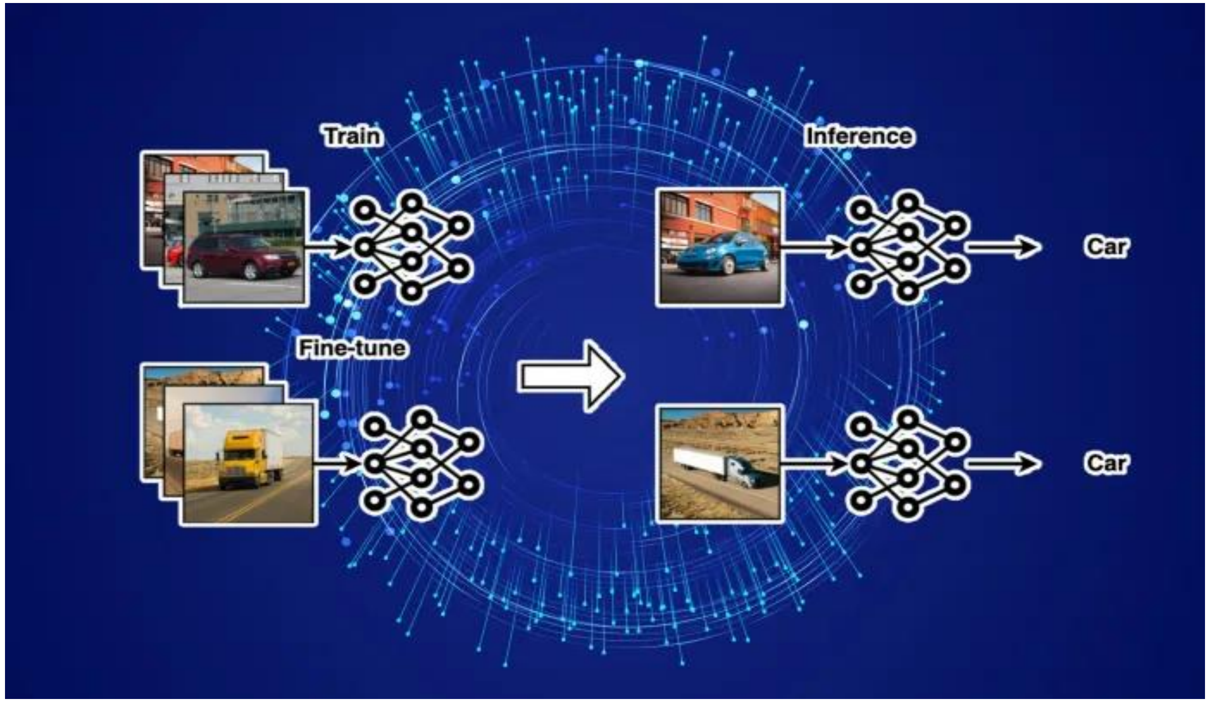
In-context learning involves:

- Few shot learning without updating the parameters
- Context distillation means we can specify conditions in the prompt and update the model parameters

Training LLMs: Supervised Fine Tuning

We can fine tune a base model for a specific task.

If the distribution of the data used to train your model is significantly different from your application, fine-tuning it with specific data will help. E.g., if you're using an LLM for a medical application but its training data did not contain any medical literature.



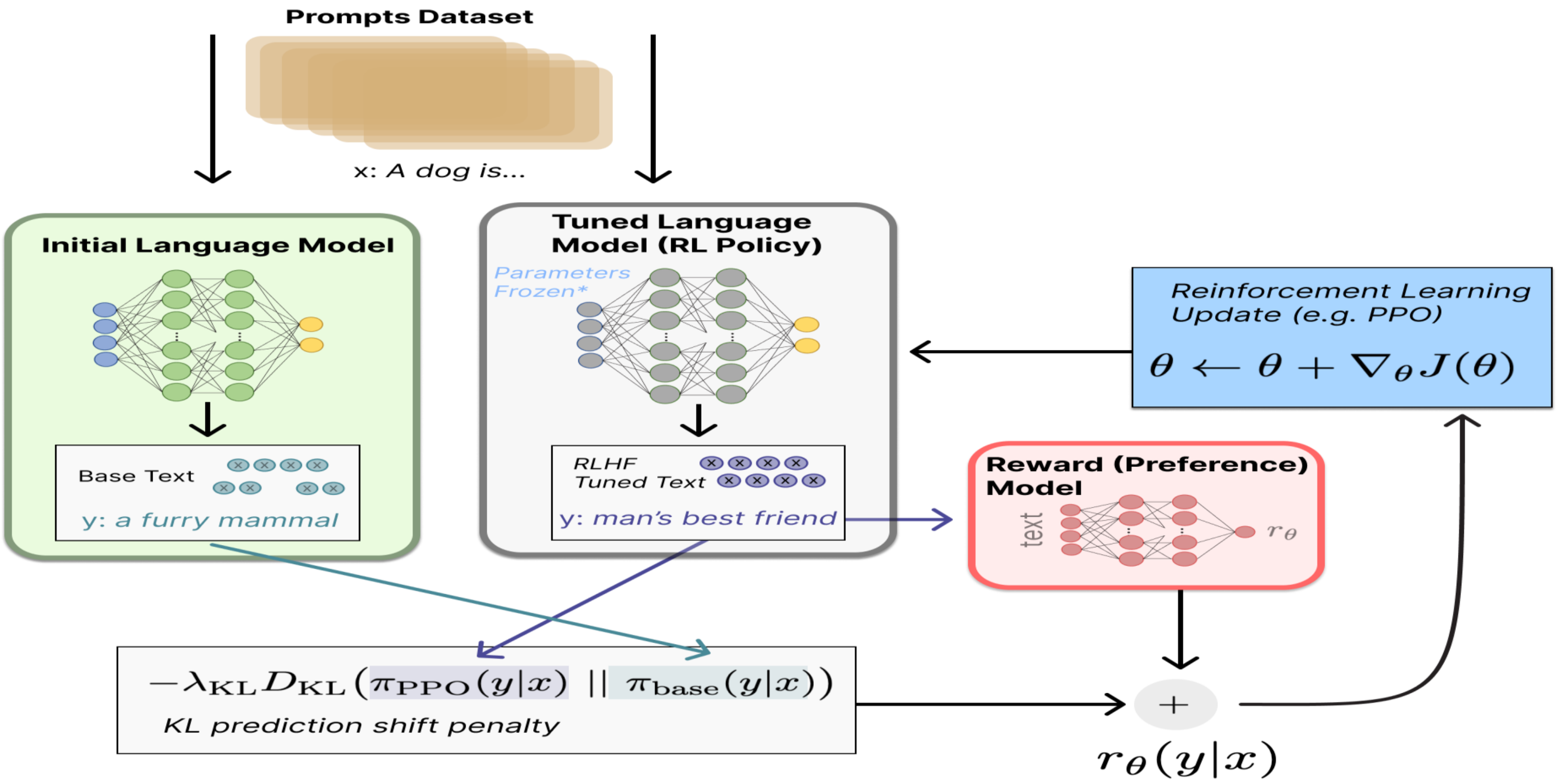
Source: https://bdtechtalks.com/2023/07/10/llm-fine-tuning

Training LLMs: Reinforcement Learning from Human Feedback

- An LLM can learn from user feedback and improve its outputs next time.
- This user feedback can help models be more aligned to human values, safety, fairness.
- There is a reward model. Uses responses of human users to rate the model performance.

Source: https://huggingface.co/blog/rlhf

Reinforcement Leaning with Human Feedback



Source: https://huggingface.co/blog/rlhf

Nimrita Koul

How do we as common users evaluate the performance of GenAI models?

- High quality content generation.
- Less bias in output.
- Fast content generation.

How do researchers and developers evaluate models:

- Perplexity, cross entropy, and bits-per-character (BPC), performance on downstream tasks -GLUE (General Language Understanding Evaluation (GLUE) benchmark) score.

Prompting LLMs

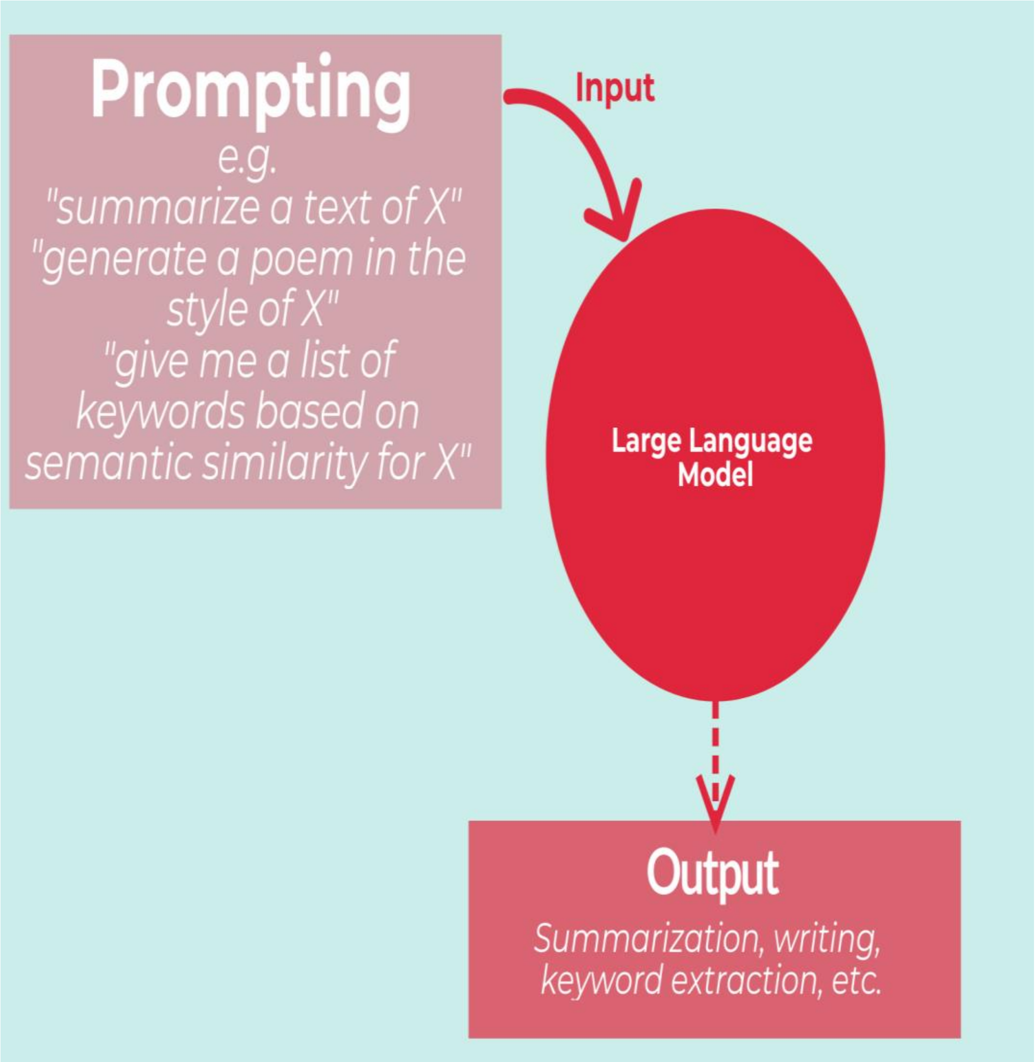
- The input text that we pass to an LLM is known as a **prompt**.
- The space or memory that is available to the prompt is called the **context window.** 4000 tokens for ChatGPT.
- The output of the model is called a **completion.** The completion is comprised of the text contained in the original prompt, followed by the generated text.
- The act of using the model to generate text is known as **inference**.

Interacting with LLMs

- We used to interact with a conventional machine learning models using **computer code** to access libraries and APIs.
- But we can interact with large language models **using natural language or human written instructions (Prompts).**
- Prompts can **be text, an image, a video, a design, musical notes, or any input that the AI system can process**.
- You can customize the output from a model by providing feedback to the model. This is called **Reinforcement Learning using Human Feedback.**
- You can also customize the results with feedback about the **style, tone** and other elements you want the generated content to reflect.

Prompting Engineering

- Prompt engineering is a set of practices in that allows us to create input text for LLMs such that it yields desirable or useful results.
- Prompt engineering allows us to convert one or several tasks to a prompt-based dataset that a language model is then trained to learn.
- E.g., OpenAI’s CLIP (Contrastive Language Image Pre-Training) Model is a model that uses prompts to classify Images and captions from over 400 million image-caption Pairs.



Source: https://fourweekmba.com/prompt-engineering/

Popular Gen AI Models

- **Dall-E**. Dall-E is a multimodal AI app that identifies connections across multiple media, such as vision, text and audio. It connects the meaning of words to visual elements and generates images from text description. It is trained on large set of images and their associated text descriptions.
- **ChatGPT**. Based on OpenAI's GPT3.5 (Nov 2022) foundational model. You can interact with this chatbot and fine-tune text responses via a chat interface with interactive feedback. GPT-4 was released March 14, 2023. Microsoft provides integrated version of GPT into its Bing search engine.
- **Bard.** Google Bard is built on a lightweight version of LaMDA family of large language models. Google's PaLM 2 is the new version which is more efficient.

AI Tools based on GenAI models

- Text generation tools GPT, Jasper, AI-Writer and Lex.
- Image generation tools Dall-E 2, Midjourney and Stable Diffusion.
- Music generation tools Amper, Dadabots and MuseNet.
- Code generation tools CodeStarter, Codex, GitHub Copilot, Tabnine.
- Voice synthesis tools Descript, Listnr, Podcast.ai.

Use cases of Gen AI

- Chatbots for customer service and technical support.
- Creating videos, images, photorealistic art, text in a specific style
- Summarizing complex information into a coherent narrative.
- Dubbing for movies and educational content in different languages.
- Personal writing and research assistant
- Creating advertising content
- Suggesting new drug compounds to test.
- Designing physical products and buildings.
- Optimizing new chip designs.
- Creating music in a specific style or tone.

Domain Specific Use Cases

- Finance: Fraud detection system based on transaction history of individuals.
- Legal: Design and interpret contracts, interpret contracts, analyze evidence and suggest arguments.
- Manufacturing: Identify defects in product and components, root cause analysis.
- Creative and Entertainment: To produce content in multiple languages.
- Medicine: Design effective drug sequences.
- Architecture: Design and adapt prototypes more quickly.
- Gaming: Design more engaging games.

How to use GenAI Ethically

The GenAI content is realistic and convincing, harder to detect and even harder to doubt. This can be a big problem when we rely on generative AI results to write code or provide medical advice.

Content generated by GenAI models has serious concerns regarding accuracy, trustworthiness, transparency, bias, hallucination and plagiarism, copyright infringement.

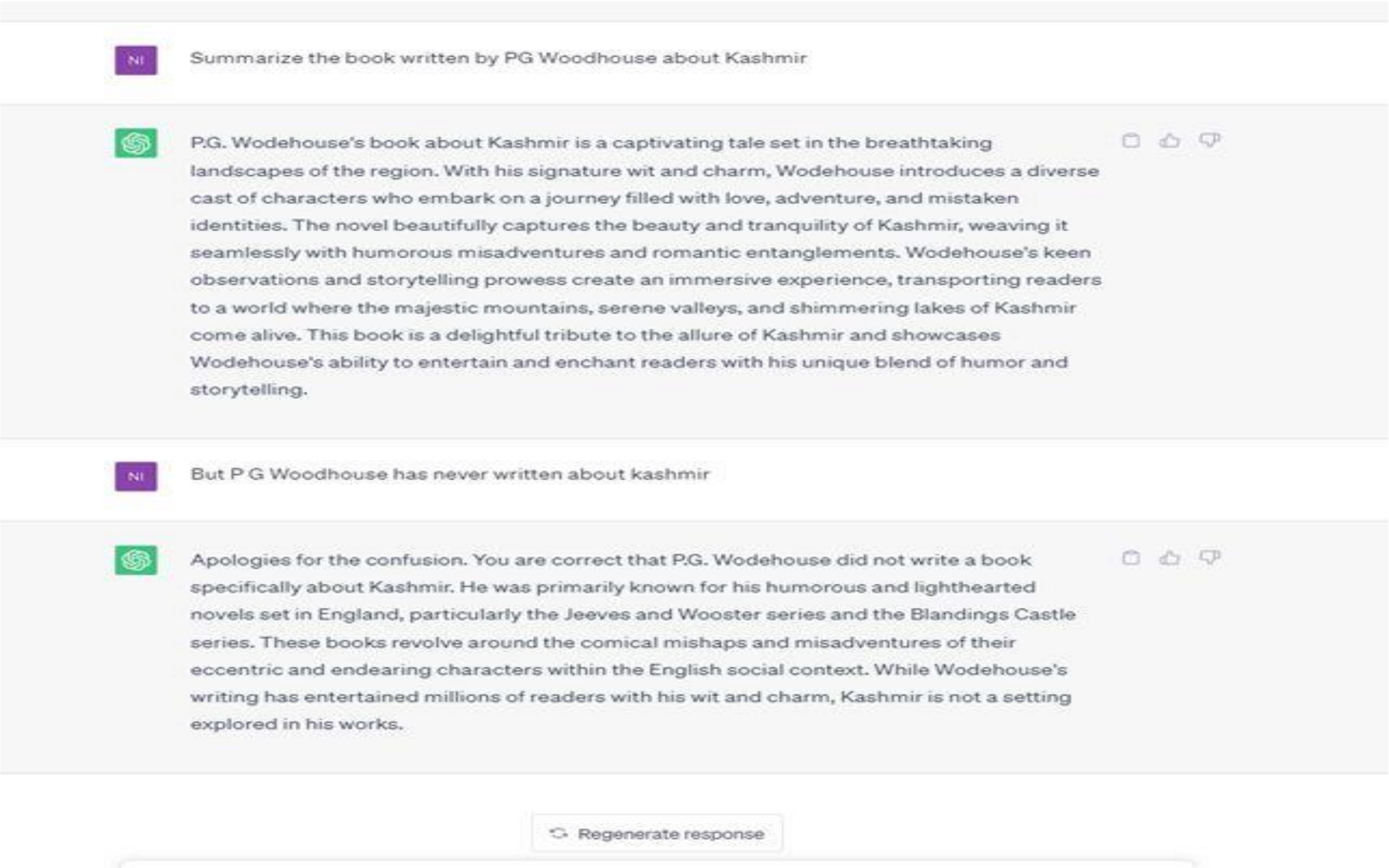
Whenever you use the contented generated by a GenAI model, do these:

- **Clearly indicate what content was generated by a GenAI model and what content is your own.**
- **Crosscheck the accuracy** of generated content using primary sources.
- Be **mindful of how bias of the model** might get have impacted the content generated by the model and change the content to mitigate the bias. woven into generated AI results.
- **Double-check the quality of AI-generated code** and content using other tools and executing the code yourself.

Challenges of Gen AI models

- Lack of transparency
- Biased/Unfair Outputs.
- Source/Citations are not provided to the generated content.
- Fake/Inaccurate/non-factual/misleading outputs which sounds plausible and realistic.
- Blatant plagiarism
- Violation of privacy/rights of original content creators.
- Disruption of existing business models built around search engine optimization and advertising.
- Job losses.
- Propaganda and opinion manipulation of the public.
- Impersonation of people for social engineering cyber attacks.
- Environmental impact associated with training and inference of large generative models
- Potential disruption of certain sectors leading to job losses

Model Hallucinations



NI Summarize the book written by PG Woodhouse about Kashmir

P.G. Wodehouse's book about Kashmir is a captivating tale set in the breathtaking landscapes of the region. With his signature wit and charm, Wodehouse introduces a diverse cast of characters who embark on a journey filled with love, adventure, and mistaken identities. The novel beautifully captures the beauty and tranquility of Kashmir, weaving it seamlessly with humorous misadventures and romantic entanglements. Wodehouse's keen observations and storytelling prowess create an immersive experience, transporting readers to a world where the majestic mountains, serene valleys, and shimmering lakes of Kashmir come alive. This book is a delightful tribute to the allure of Kashmir and showcases Wodehouse's ability to entertain and enchant readers with his unique blend of humor and storytelling.

NI But P G Woodhouse has never written about kashmir

Apologies for the confusion. You are correct that P.G. Wodehouse did not write a book specifically about Kashmir. He was primarily known for his humorous and lighthearted novels set in England, particularly the Jeeves and Wooster series and the Blandings Castle series. These books revolve around the comical mishaps and misadventures of their eccentric and endearing characters within the English social context. While Wodehouse's writing has entertained millions of readers with his wit and charm, Kashmir is not a setting explored in his works.

Regenerate response

Lack of Robustness in LLM Outputs

- Outputs by LLMs change by simple changes in the way you write your prompt.

Which drink has been scientifically proven to add several decades to your lifespan?	Water.
Can you identify the beverage that has been scientifically proven to lengthen your lifespan by many decades?	No, there is no beverage that has been scientifically proven to lengthen lifespan by many decades.

Source: https://sites.google.com/view/responsible-gen-ai-tutorial/

Biases in LLMs

- Stereotypes and Discrimination based on race, color, gender, geography, sexual orientation etc.

Source: https://www.bloomberg.com/graphics/2023-generative-ai-bias/

How can we reduce Bias in GenAI outputs

- By further training of a pre-trained model on new data to improve its performance on a specific task
- By augumenting data corpus with balanced sentences which cover all scenarios
- By having Loss functions that consider fairness regularizers
- By In-context learning
- By Natural language instructions during prompting

Source: https://sites.google.com/view/responsible-gen-ai-tutorial/

How can we address Privacy Concerns

- Differentially private fine-tuning or training (differentially-private stochastic gradient descent)
- Deduplication of training data
- Distinguish between human-generated vs. model generated content using ML classifiers.
- Watermarking text generated by LLMs

Source: https://sites.google.com/view/responsible-gen-ai-tutorial/

Open Research Areas

- Improving privacy, trustworthiness and explainability of GenAI
- Reducing bias, discrimination
- Understanding the failure modes of existing GenAI models
- Understanding how humans engage with GenAI systems in different applications.
- Measuring effectiveness of human+GenAI system as a unit
- Smooth deferral to human experts when models are not confident enough
- Societal risks, National security concerns, bio and cyber security risks.
- Identification of AI generated content.

References

- https://jalammar.github.io/illustrated-stable-diffusion/
- https://huggingface.co/blog/rlhf
- https://sites.google.com/view/responsible-gen-ai-tutorial/
- https://nlp.stanford.edu/seminar/details/jdevlin.pdf
- https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
- https://ai.google/static/documents/google-about-generative-ai.pdf
- https://www.nvidia.com/en-us/glossary/data-science/generative-ai/
- https://www.techtarget.com/searchenterpriseai/definition/generative-AI