COMP9444: Neural Networks and Deep Learning

Generative AI Applications

Jingying Gao
School of Computer Science and Engineering
July 31, 2024

What cognitive capabilities do humans have?

Inputs

- See – Visual Processing
- Listen -Auditory Processing
- Read – Textual Understanding
- Feel - Sensory and Emotional Processing

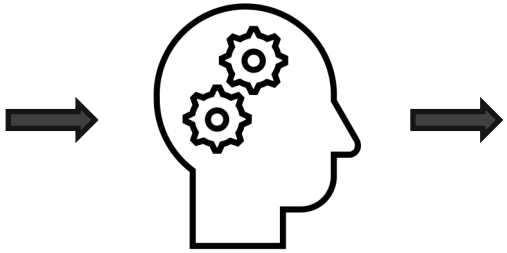
Outputs

- Speaking/Singing – Verbal Expression
- Writing – Textual Expression
- Painting – Visual Artistic Expression
- Thinking – Cognitive Reasoning

What capabilities does AI have?

Inputs

- ❑ See – Image / Video
- ❑ Listen – Audio / Language
- ❑ Read – Text
- ❑ Feel – Sensor



Outputs

- ❑ Speaking / Singing
- ❑ Writing
- ❑ Painting
- ❑ Thinking - Question Answering & Reasoning

Multimodal AI

Generative AI

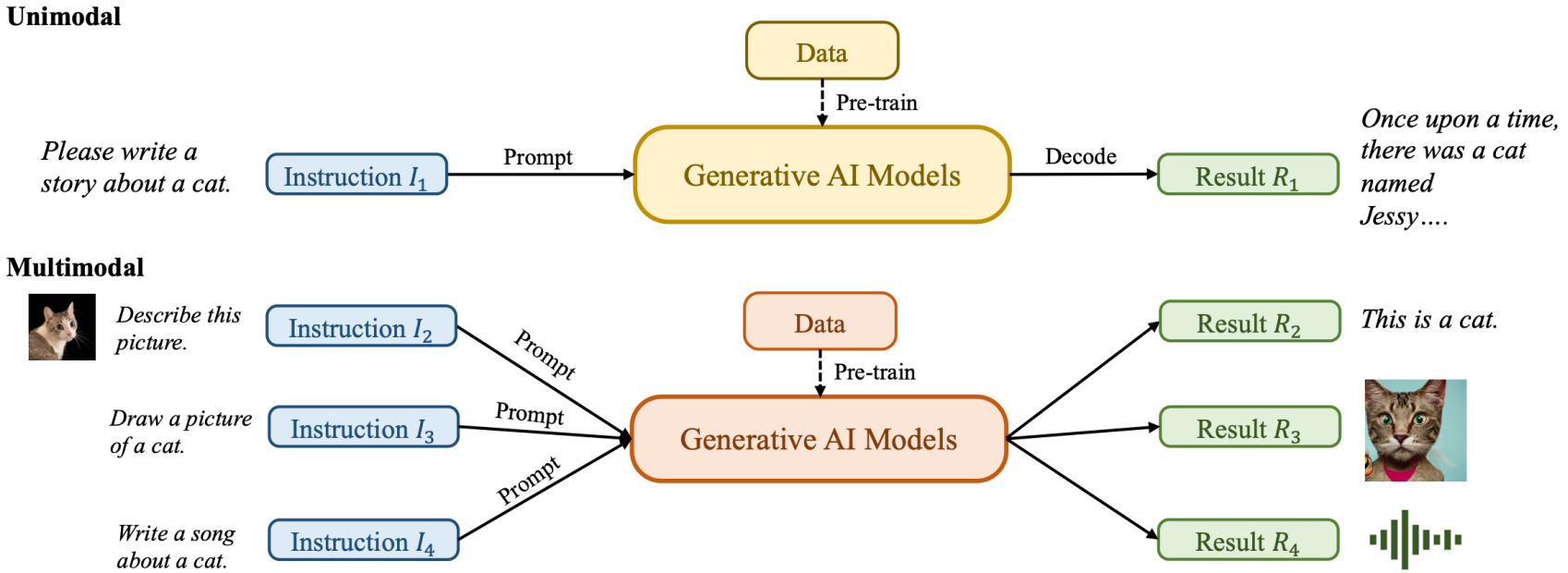
Generative AI

- What Is Generative AI?
- What Are the Different Types of Generative AI Tasks?
- How Does Generative AI Work?
- How to Build an Interactive App Using Generative AI?

What is Generative AI?

- Generative AI (GenAI) is artificial intelligence capable of generating text, images, videos, or other data using generative models, often in response to prompts.
- Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.

What Are the Different Types of Generative AI Tasks?



Source: https://arxiv.org/pdf/2303.04226.pdf

Generative AI Tasks

- Text Generation: Text-to-Text, Image-to-Text, Speech-to-Text
- Image Generation： Image-to-Image, Text-to-Image, Talking face
- Audio Generation: Text-to-Audio, Image-to-Audio
- Video Generation: Text-to-Video, Image-to-Video

Demo of Generative AI Tasks

- Generate a Painting
- Image to Video
- Outfit Anyone
- Generate a Song

Demo of Generative AI Tasks

- Generate a painting
- https://huggingface.co/spaces/multimodalart/stable-cascade
- Image to Video
- https://huggingface.co/spaces/multimodalart/stable-video-diffusion
- Generate a song
- https://suno.com/

Generate a Song

Prompt
"Welcome, students, to the AI course. Today's lecture will cover generative AI and multimodal AI. We are going to explore various topics, including stable diffusion models and LLM models, among others."

Enjoy the below songs:

Future of Creation

Rise Above

Which one is real?

Which one is real?

Which one is real?

Generative AI

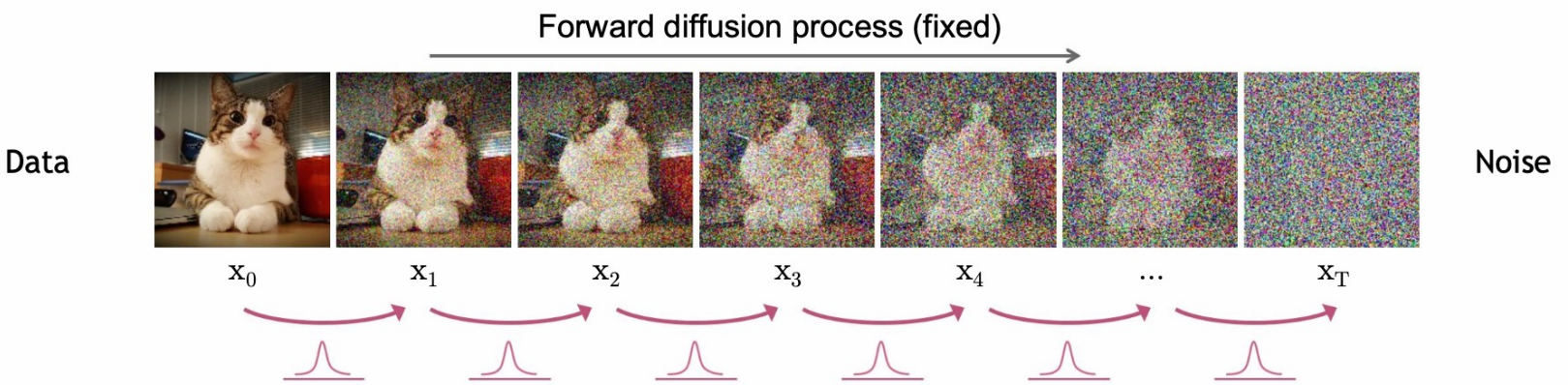
- How do Text-to-Image Generation Models Work?
- How does the Diffusion Model Work?
- Fine-tune a Diffusion Model with a Customize Dataset.
- How to Build an Application based on Diffusion AI Models?

How Did the Idea Come About?

The central idea behind diffusion models is inspired by the thermodynamics of gas molecules, whereby the molecules diffuse from areas of high density to low density.

This movement is often referred to in physics literature as the increase of entropy or heat death.

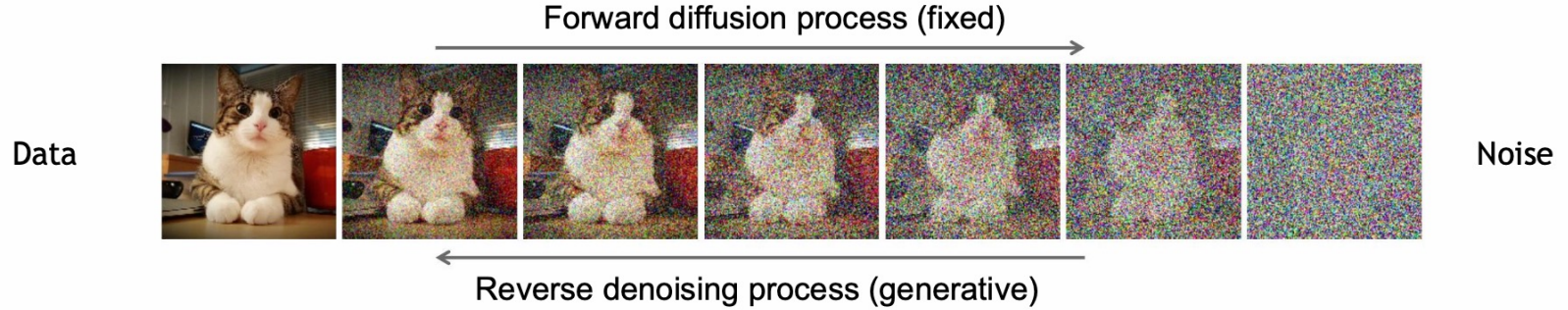
How Does an AI Model Generate an Image? (Stable Diffusion/DALL-E)



Diffusion models work by gradually adding Gaussian noise through a series of T steps into the original image, a process known as diffusion.

Source: Diffusion, Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models

How does AI generate an Image? (Stable Diffusion / Dall-E)

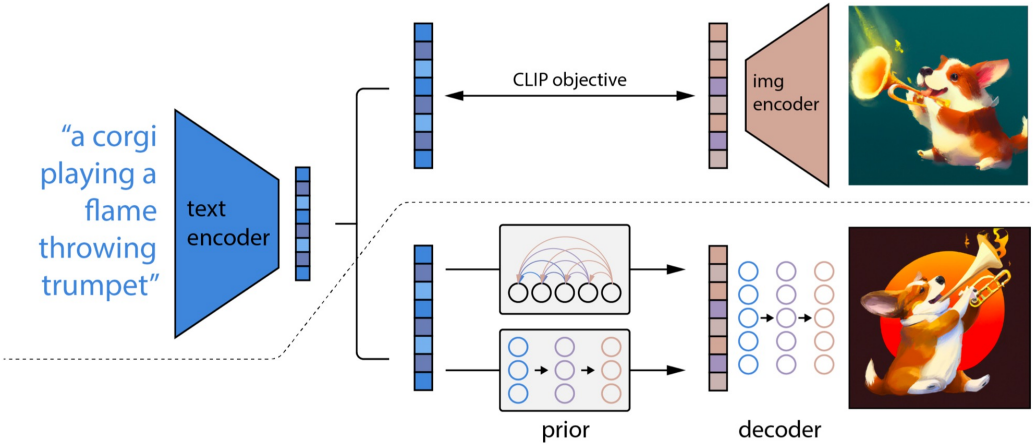


- Backward Diffusion Process.
- We take a neural network Unet and learn to reverse this diffusion process.
- The reverse diffusion process, learn from the last right noise image, generate image from step t to step t-1.
- From step t, use a neural network Unet to predict noise, then the predicted noise will be subtracted from the image.

Source: Diffusion, Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models

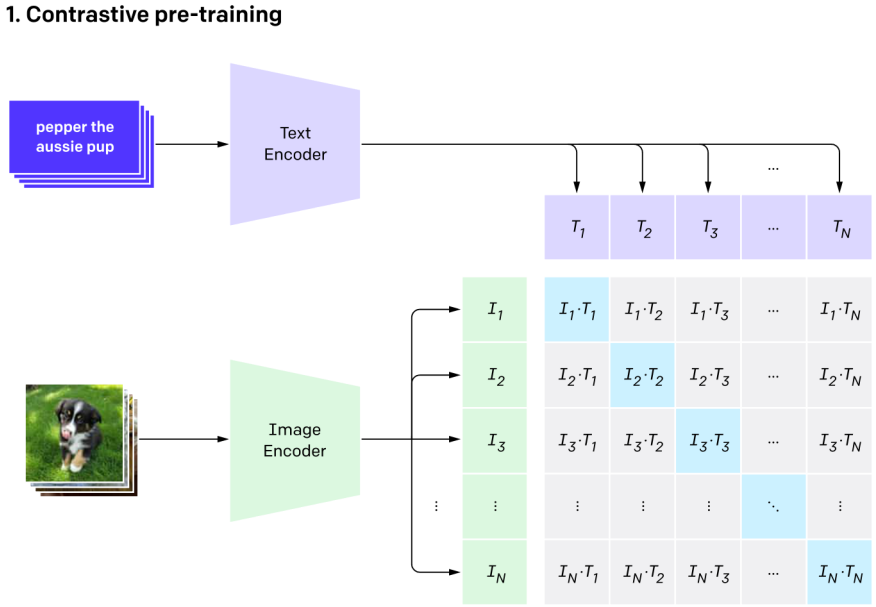
How to generate images that correspond with a piece of text?

- We've learned how diffusion can help us to generate an image from random noise, but the model is only able to generate random images.
- How can we make use of this model to synthesize images that correspond with a class name in our training data, a piece of text, or another images?



Source: Radford et al., Learning Transferable Visual Models From Natural Language Supervision

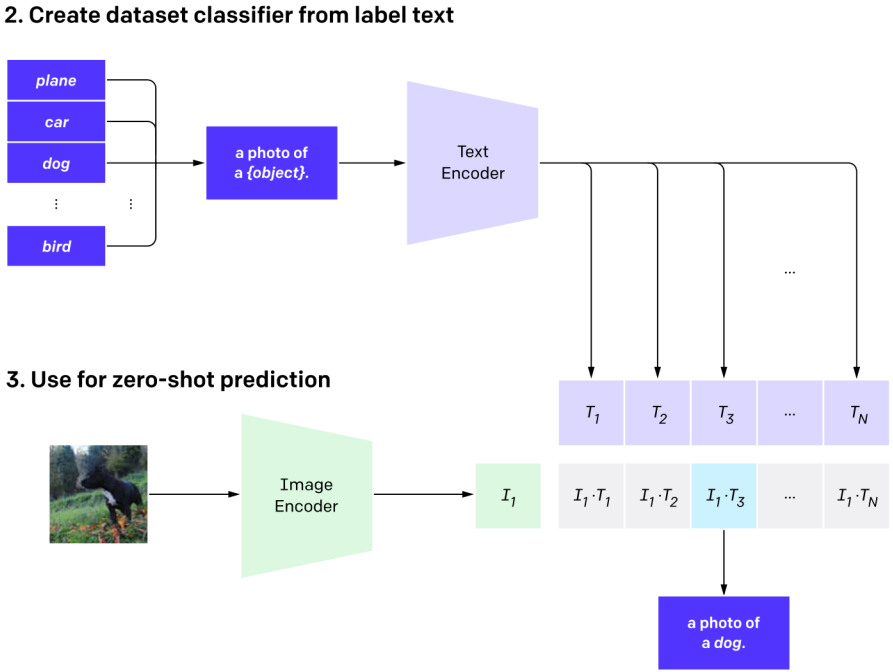
What is Contrastive Learning?



- Convert Image to Vector via Image Encoder.
- Convert Text to Vector via Text Encoder.
- Learn the pairs of image vector and text vector. The diagonal is the positive samples, others are negative samples.
- Minimize the negative samples.

Source: Radford et al., Learning Transferable Visual Models From Natural Language Supervision

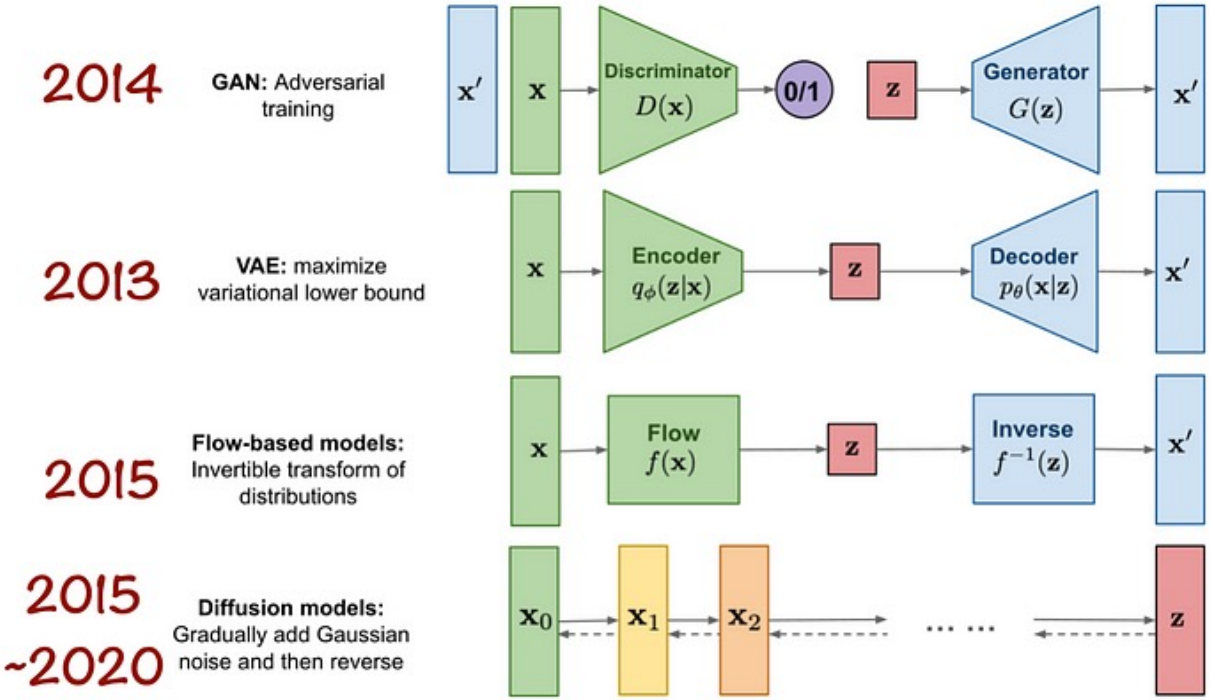
How to generate images that correspond with a piece of text?



- CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset.
- We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.
- Contrastive loss, is used to map vectors that model the similarity of input items.

Source: Radford et al., Learning Transferable Visual Models From Natural Language Supervision

Overview of different types of generative models



Source: What are Diffusion models?

GAN Adversarial Training

- **Discriminator**: It is a classifier that tries to distinguish between real data (from the training dataset) and fake data created by the generator. It's trained to accurately classify data as real or fake.
- **Generator**: It generates new data instances that resemble the training data. It starts with random noise and gradually learns to produce data (like images) that look similar to the actual data it's been trained on. Its goal is to create data so convincing that the discriminator can't tell it apart from real data.

VAE - maximize variational lower bound

A Variational Autoencoder (VAE) aims to learn a representation of input data in a latent space by maximizing the variational lower bound.

- **Encoder**: It takes the input data and transforms it into a distribution over the latent space. The encoder outputs parameters to this latent distribution, typically the mean and variance, effectively compressing the input data into a compact representation.
- **Decoder**: Starting from the latent space representation, the decoder attempts to reconstruct the original input data from the latent distribution. It learns to map the latent variables back to the original data space, aiming to minimize the difference between the original data and its reconstruction.

How to fine-tune the stable diffusion model?

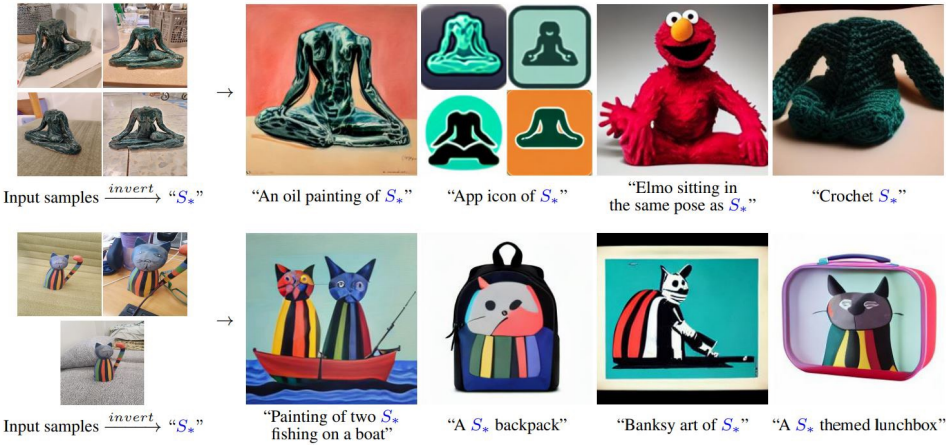
- Why do we need fine-tune the stable diffusion model?
- How to fine-tune the stable diffusion model?

Why do we need to fine-tune the stable diffusion model?

- Stable Diffusion was trained on 2.3 billion English image-text pairs. So this means that there are 2.3 billion pictures and their captions that are matched up, and this is the data that is used in training the Stable Diffusion model.
- Does it include UNSW Style images or if we want to generate some UNSW particular classes, such as UNSW library, UNSW campus?

How do we fine-tune the diffusion model to generate style images?

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion



We learn to generate specific concepts, like personal objects or artistic styles, by describing them using new "words" in the embedding space of pre-trained text-to-image models. These can be used in new sentences, just like any other word.

The work builds on the publicly available Latent Diffusion Models

How do we fine-tune the diffusion model to generate style images?

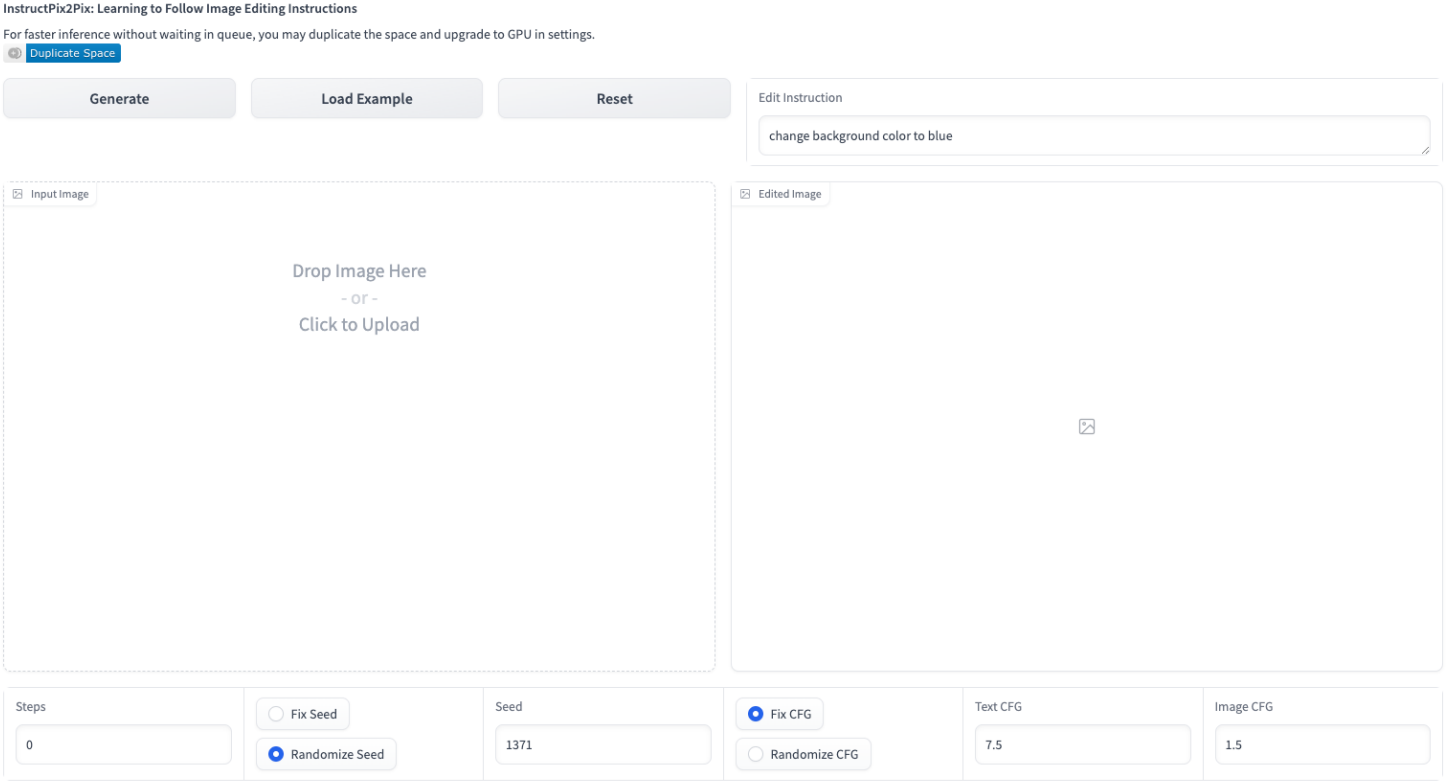
- Prepare Customized Dataset: Image and Text Pairs.
- Download pre-trained Diffusion model's weights.
- Define hyper-parameters for the training process.
- Training.
- Run for generating images.

Demo Code Fine-tune model

https://colab.research.google.com/drive/1F9UqlsHHmZZvgqAw78VqbAYWlTwep9Qv?usp=sharing

Fine-tune diffusion model

How to build an interactive diffusion application?



How to build an AI image generation application?

- Huggingface models
- Huggingface Space
- Gradio

Gradio

- Gradio is a Python library that allows you to quickly create customizable web apps for your machine learning models and data processing pipelines.
- Gradio apps can be deployed on Hugging Face Spaces for free.

Demo Instruct Pix-2-Pix

Demo Instruct Pix 2 Pix

https://huggingface.co/spaces/timbrooks/instruct-pix2pix

https://huggingface.co/spaces/timbrooks/instruct-pix2pix

A Taste of Code

```
import random
import gradio as gr
import torch
from PIL import Image
from diffusers import StableDiffusionInstructPix2PixPipeline

example_instructions = [
    "Change the hair color to blue",
    "Add a mustache",
    "Change the background to a beach",
    "Add sunglasses",
]

model_id = "timbrooks/instruct-pix2pix"
pipe = StableDiffusionInstructPix2PixPipeline.from_pretrained(model_id, torch_dtype=torch.float16, safety_checker=None).to("cuda")

def generate_image(input_image: Image.Image, instruction: str):
    edited_image = pipe(instruction, input_image).images[0]
    return edited_image

iface = gr.Interface(
    fn=generate_image,
    inputs=[
        gr.Image(type="pil", label="Take a picture", sources=["webcam"]),
        gr.Dropdown(choices=example_instructions, label="Edit Instruction"),
    ],
    outputs=gr.Image(type="pil", label="Edited Image"),
    title="InstructPix2Pix",
    description="Capture an image from the webcam, enter a prompt, and change the image.",
)

iface.launch(debug=True)
```

Generative AI - LLM

- What is LLM?
- How does GPT work?
- How to use VectorDB to build a customize chatbot?

What is LLM?

- A large language model (LLM) is a language model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification.
- LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process.
- LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.

What does GPT stand for?

GPT =

Generative Pre-trained Transformer

What does GPT stand for?

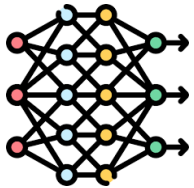
Generative

Generate New Text

“AI course is a general course to learn Artificial Intelligence.”

Pre-trained

Pretrained refers to how the model went through a process of learning from massive amount of data.

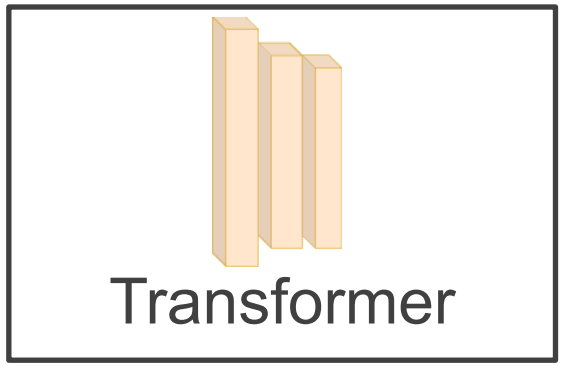


Transformer

A deep neural network model that transforms or changes an input sequence into an output sequence by using attention.

How does Transformer work?

Attention is all you need



注意力就是你所需要的

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

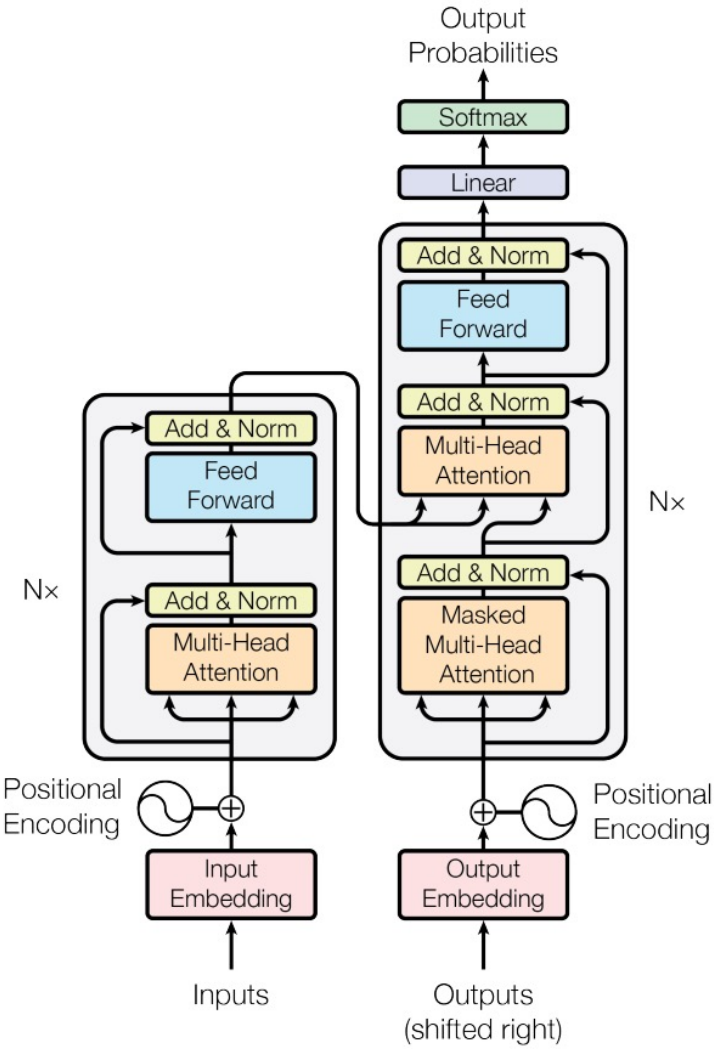
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

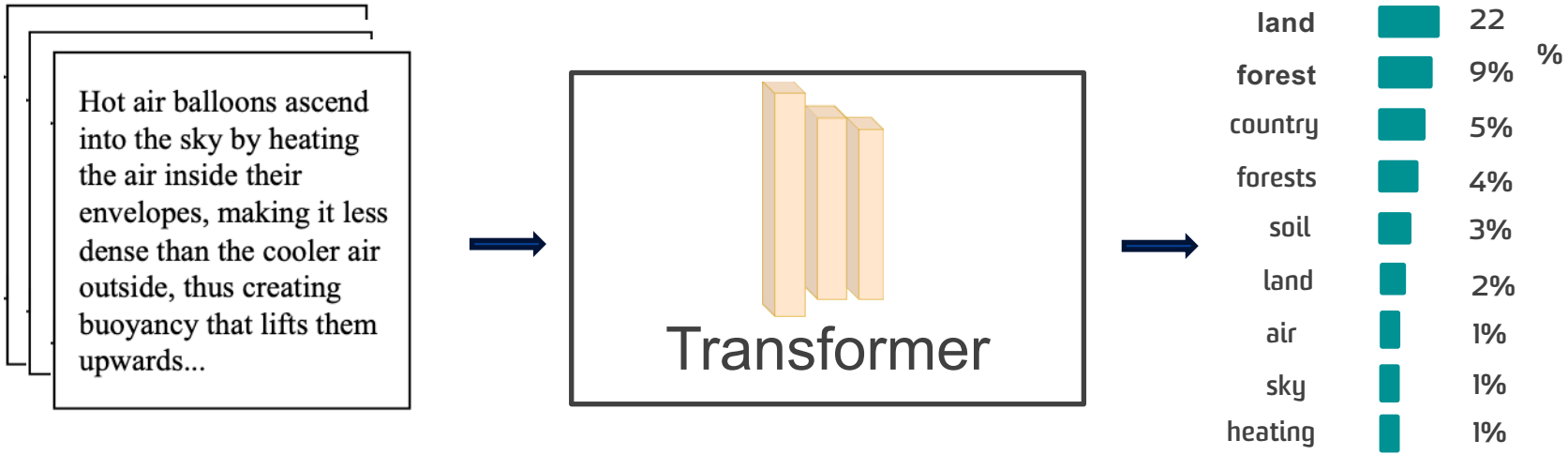
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to

Transformer Architecture



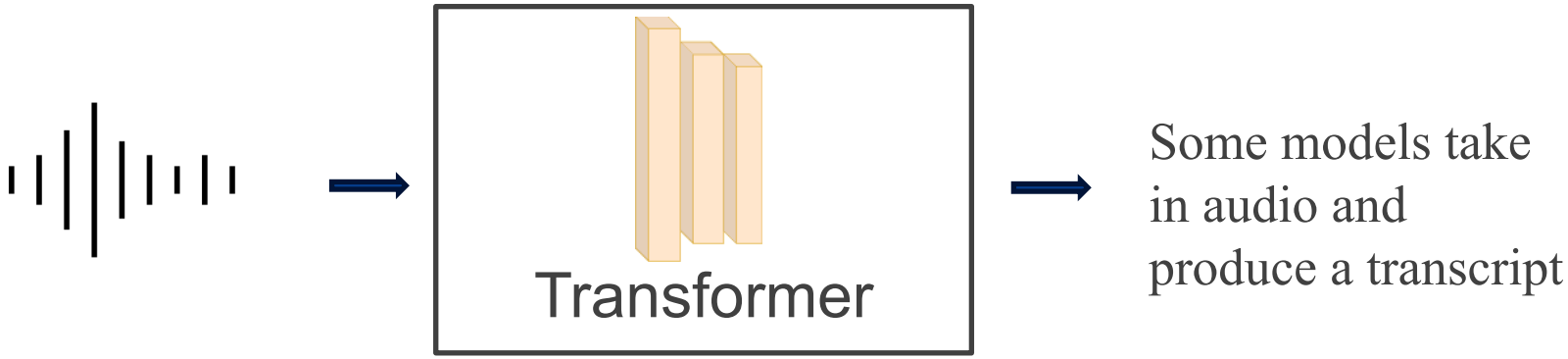
Source: Vaswani et al., Attention Is All You Need

How does GPT work?



- GPT model that is trained to take in a piece of text, or image, sound, to produce a prediction for what comes next in the passage.
- That prediction takes the form of a probability distribution over many different chunks of text that might follow.

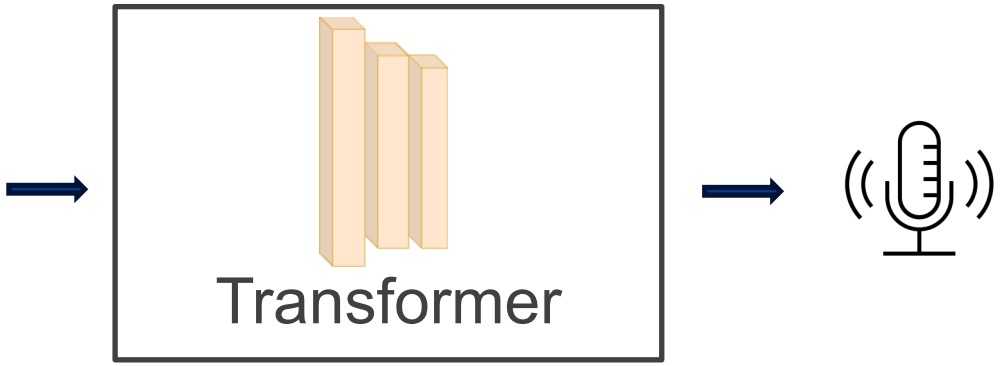
Models: Voice-to-Text



- The model takes in audio and produce a transcript.
- Example: OpenAI whisper model.

Generative AI Model: Text-to-Voice

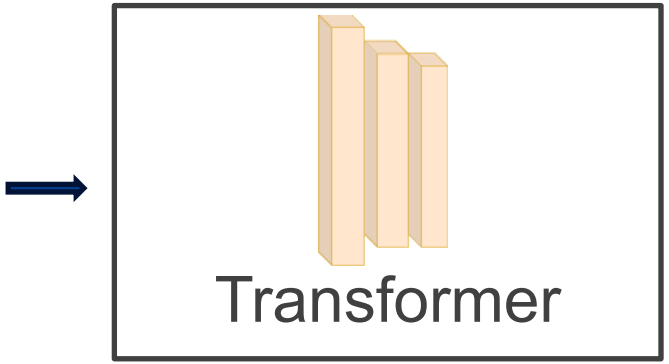
This is an AI course model example, producing synthetic speech from text.



- Transformer produces synthetic speech from text.

Generative AI Model: Text-to-Voice

Generate a cute cat is studying AI course.



Generated by ChatGPT

- The model takes in a text description and produce an image are based on Transformers.

LLM Applications

- LangChain
- RAG
- Agents

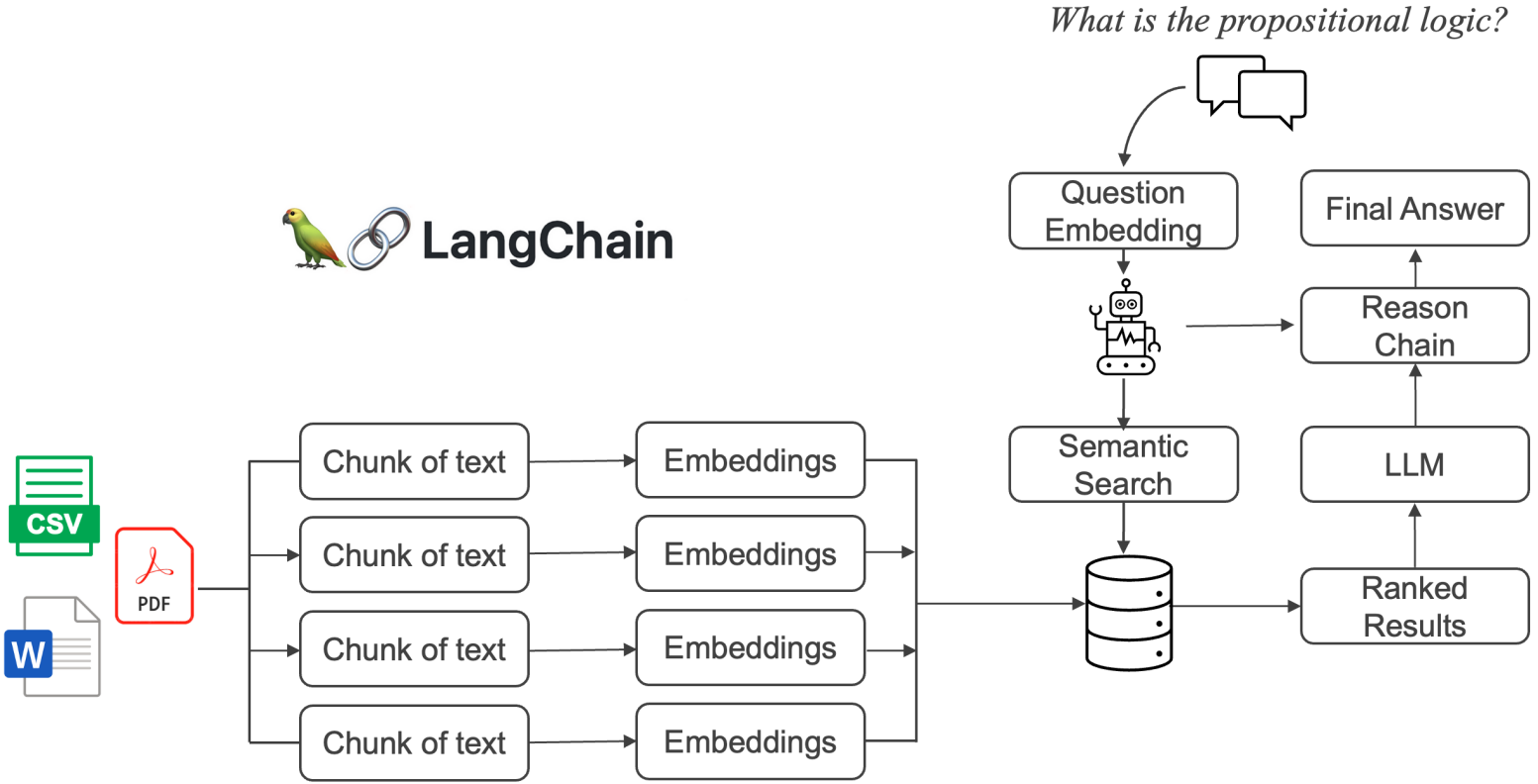
What is LangChain?

LangChain is an open-source framework for the development of applications that use large language models, such as GPT-4.

Lang = Language which refers to the use of large language models like GPT4.

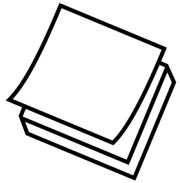
Chain = Chaining that connect/chain these language models together to build applications

What is LangChain Architecture?



What is RAG?

RAG is when the LLM looks for additional information, augments the query and generate an answer with new information.

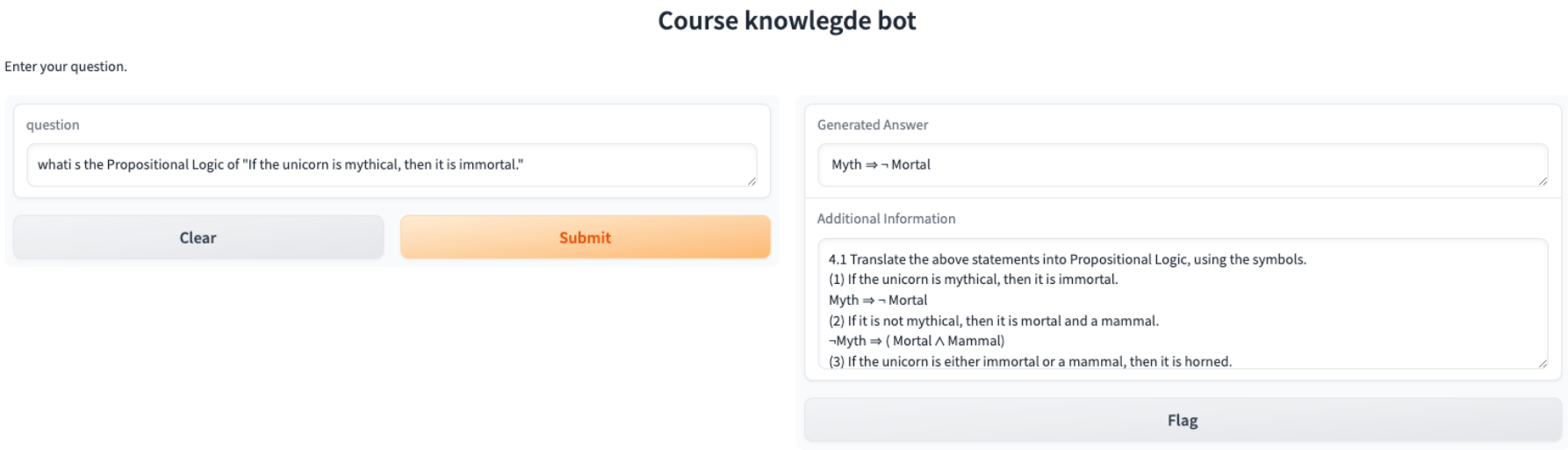


Retrieval: LLM finds extra information from the knowledge base documents based on “similarity” metric

Augmentation: LLM integrates the extra information into the prompt

Generation: LLM generates an answer using the augmented prompt

How to build an AI Tutorial Knowledge Bot?



LLM Shortcomings

- Most time, we use LLMs in zero-shot mode, prompting a model to generate final output token by token without revising its work.
- This is akin to asking someone to compose an essay from start to finish, typing straight through with no backspacing allowed, and expecting a high-quality result.

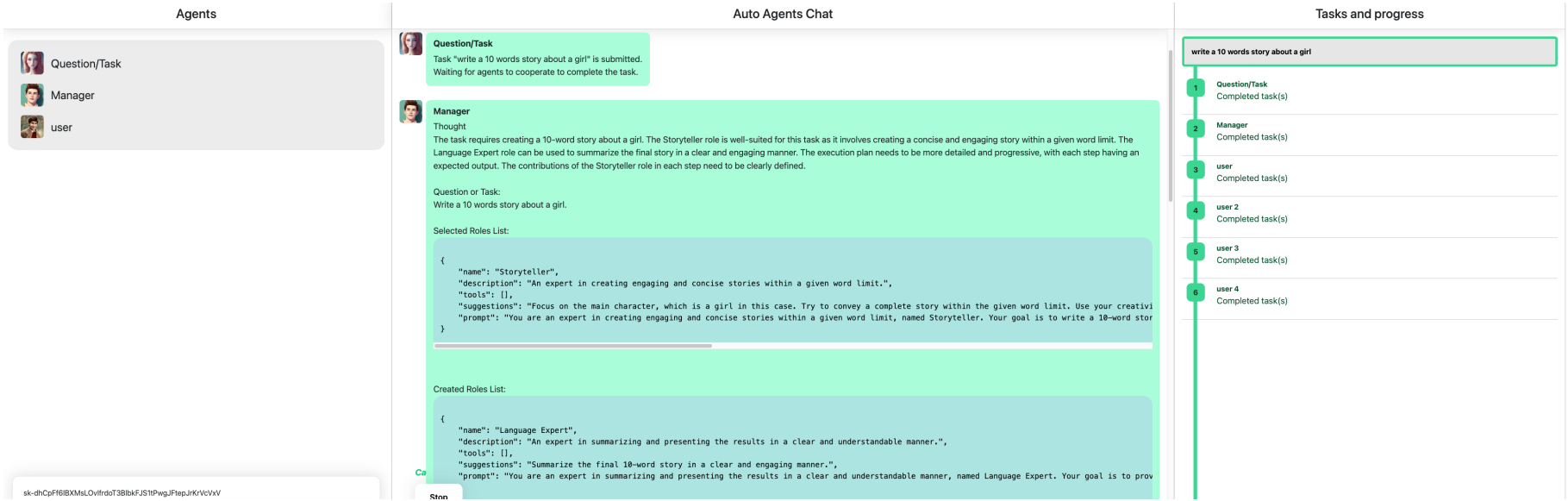
LLM Agents

- The experience of prompting ChatGPT, receiving unsatisfactory output, delivering critical feedback to help the LLM improve its response, and then getting a better response.
- What if we automate the step of delivering critical feedback, so the model automatically criticizes its own output and improves its response?
- LLM Agents allows LLM to iterate over documents many times, take self-reflection and improve the results.

LLM Agents - Example

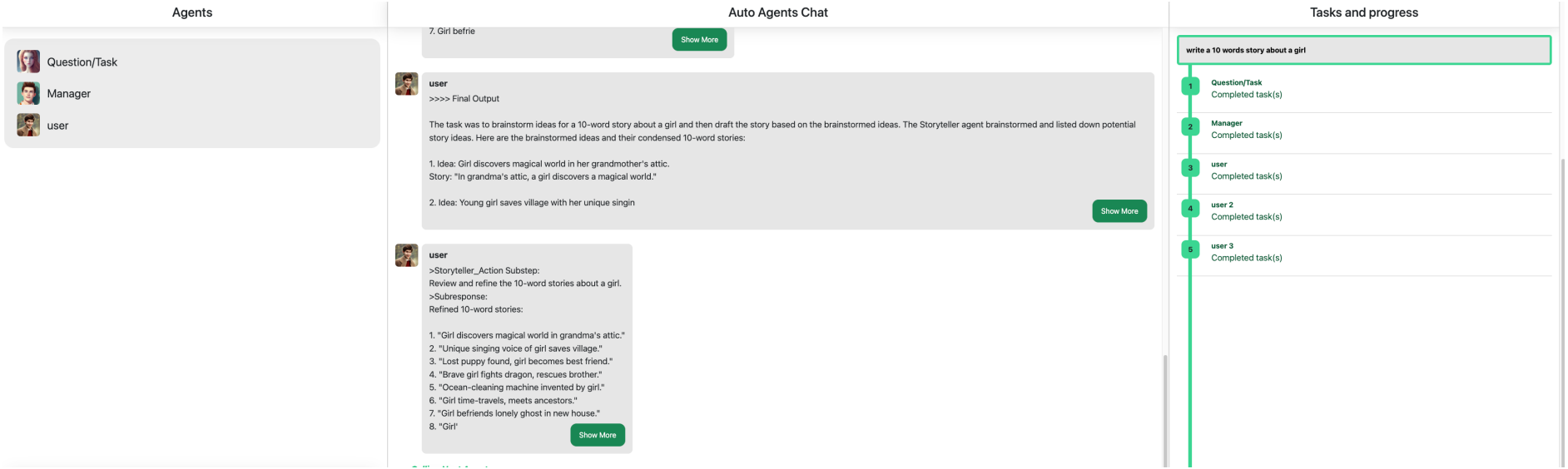
- We use a system built with Reflection to write some ABC code.
- Then, the AI will take this code, along with a prompt like “check the correctness of this code and tell me how to modify it,” and return it to the AI.
- The AI might identify bugs within it, and through this process, the AI completes its self-iteration.
- Although the quality of the revised code can't be guaranteed, the results are generally better on the whole.

LLM Agents - Demo



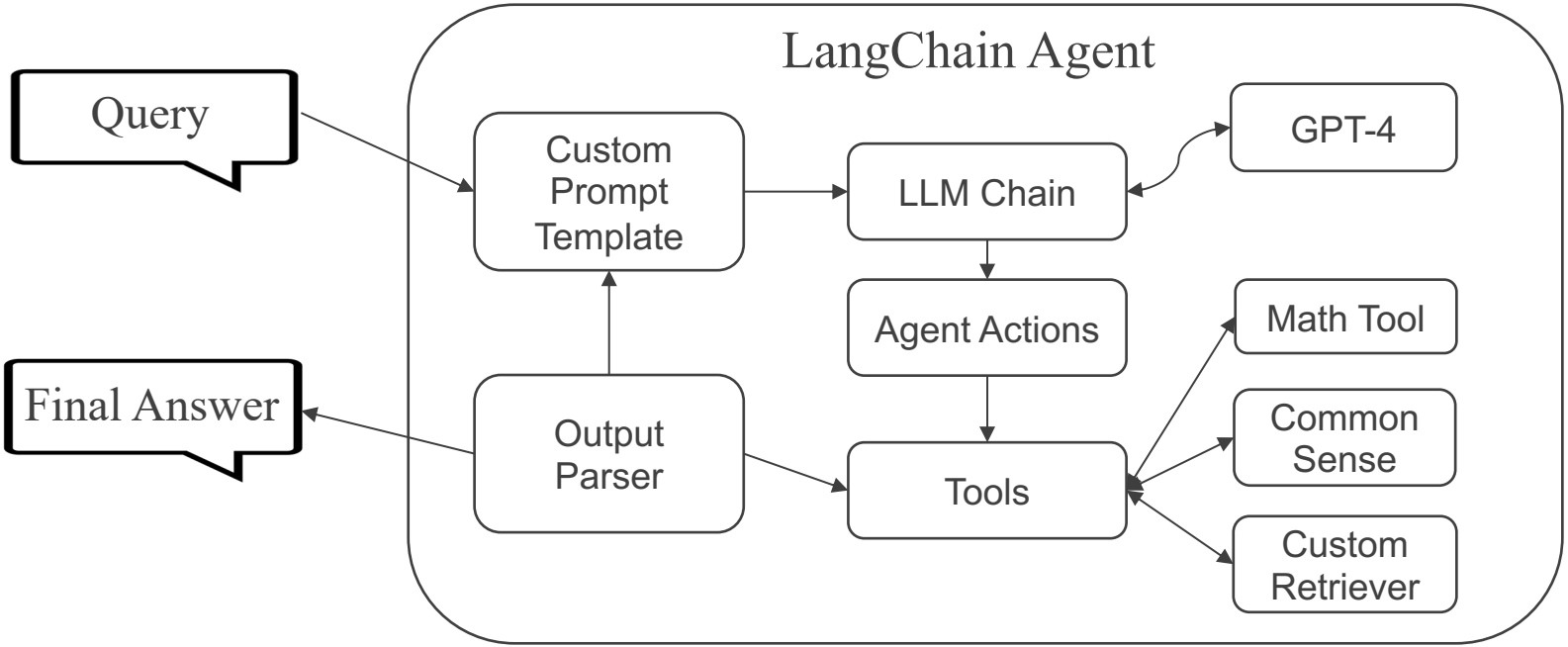
Source: [Auto Agents]

LLM Agents - Demo

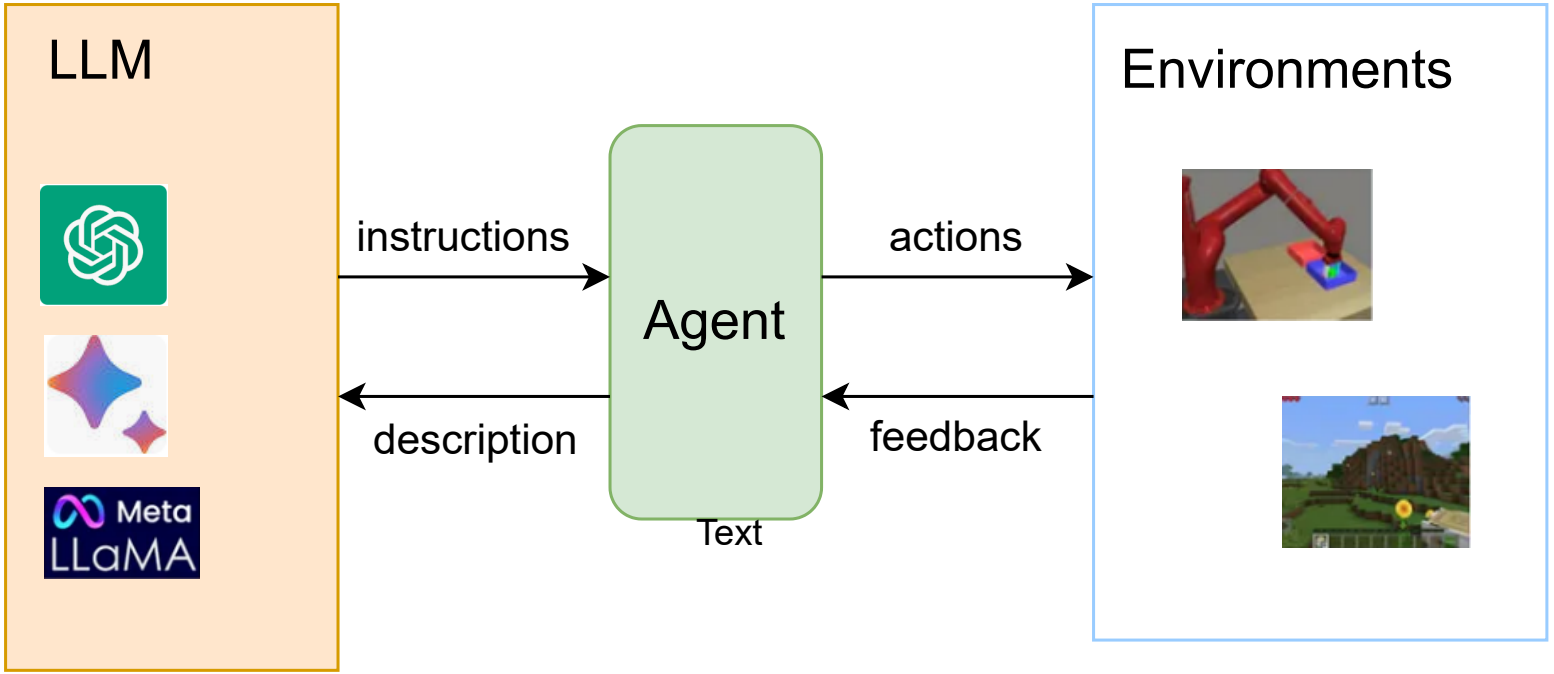


Source: Auto Agents

What is LLM Agent?

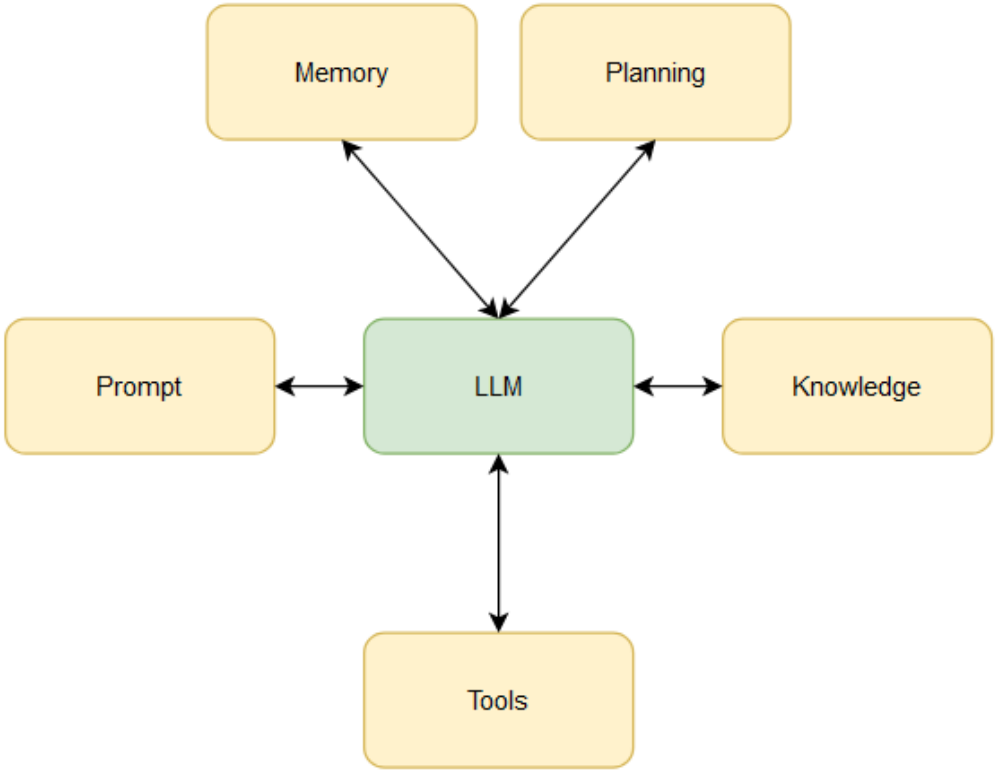


LLM Agents – General Framework



Source: What is an LLM Agent and how does it work?

Components of LLM Agents



Source: What is an LLM Agent and how does it work?

LLM Agents

https://huggingface.co/spaces/LinkSoul/AutoAgents

Future Directions

- Explainable AI
- Multimodal Reasoning
- Logical Reasoning in Multimodal AI Tasks
- AI Ethics