CA6001 Chapter 6
Generative AI

Dr Zhang Jiehuang

College of Computing and Data Science

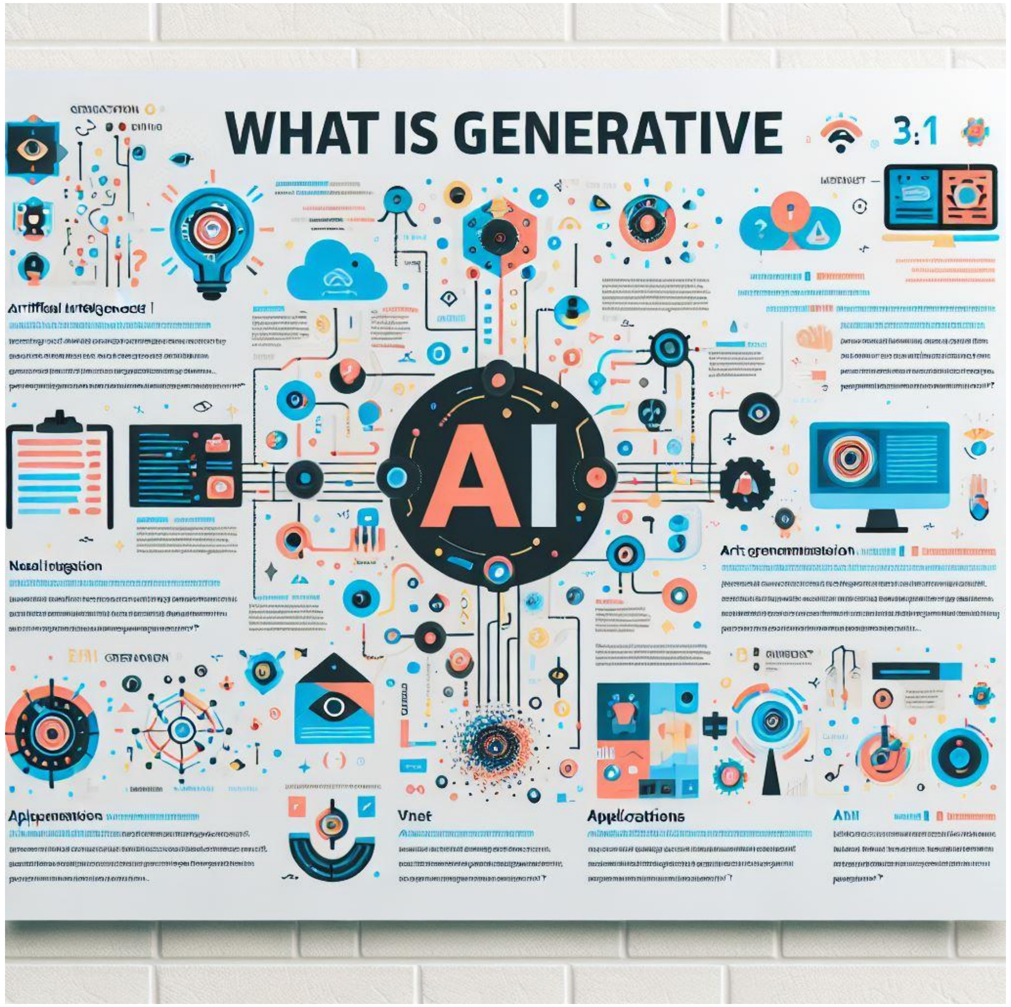
Nanyang Technological University

email: *jiehuang.zhang@ntu.edu.sg*

Chapter 6 – Generative AI (GenAI)

1. What Is Generative AI
2. Use Cases: OpenAI Sora
3. Generative Adversarial Networks
4. Variational AutoEncoders
5. Diffusion Models
6. Transformers Revisited
7. Training Large Networks
8. Ethics and Regulation

What is Generative AI?



Generative AI learns the underlying, finds patterns and structures from data. Then uses it to create new data

Generative AI boomed in the early 2020s due to transformer based deep learning, and we are still amid the boom, with new technologies emerging everyday

LLMs – ChatGPT, Copilot, Gemini

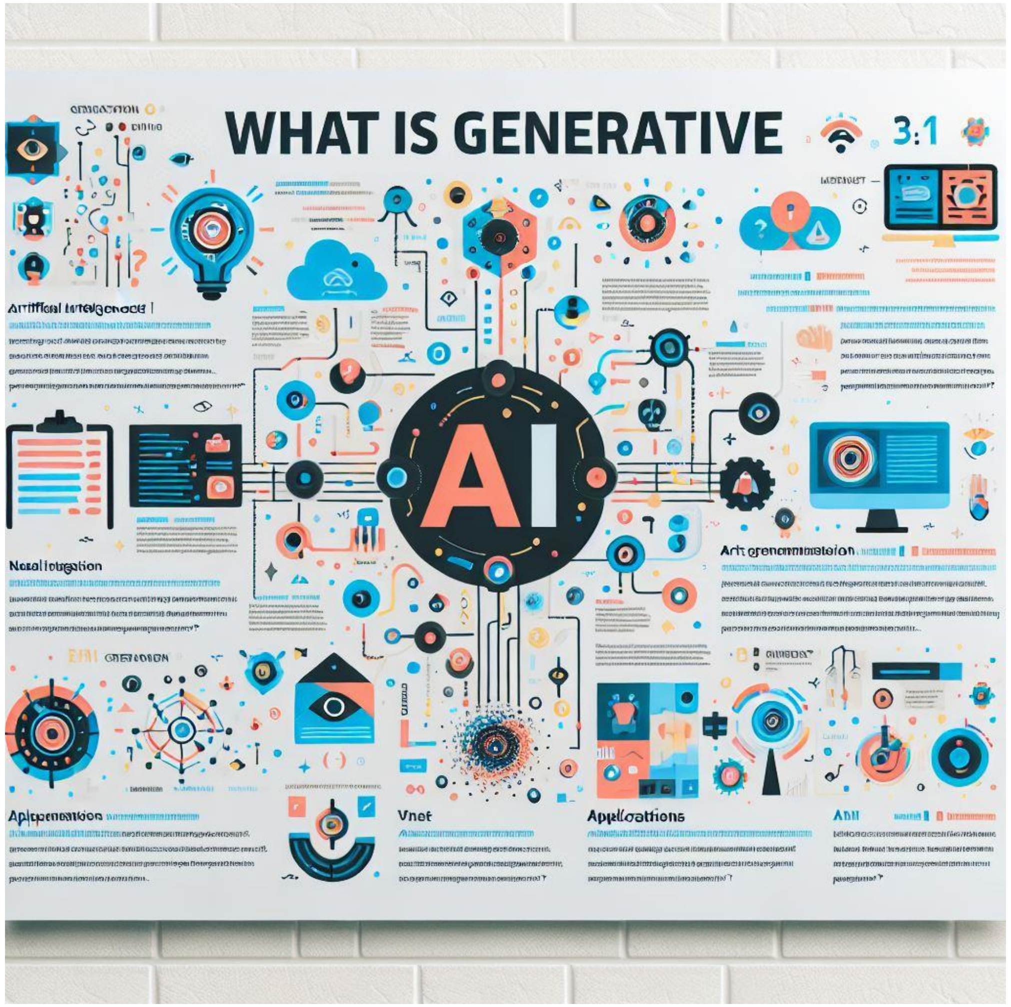
Image – Stable Diffuson, Midjourney, DALL-E

Music – ChucK, Jukedeck, MorpheuS

Video – SORA, RunwayML, Make A Video

3D Spatial Models – Computer Aided Design

What is Generative AI?



Generative AI can be used to do data augmentation, creating new data in some areas where data is limited

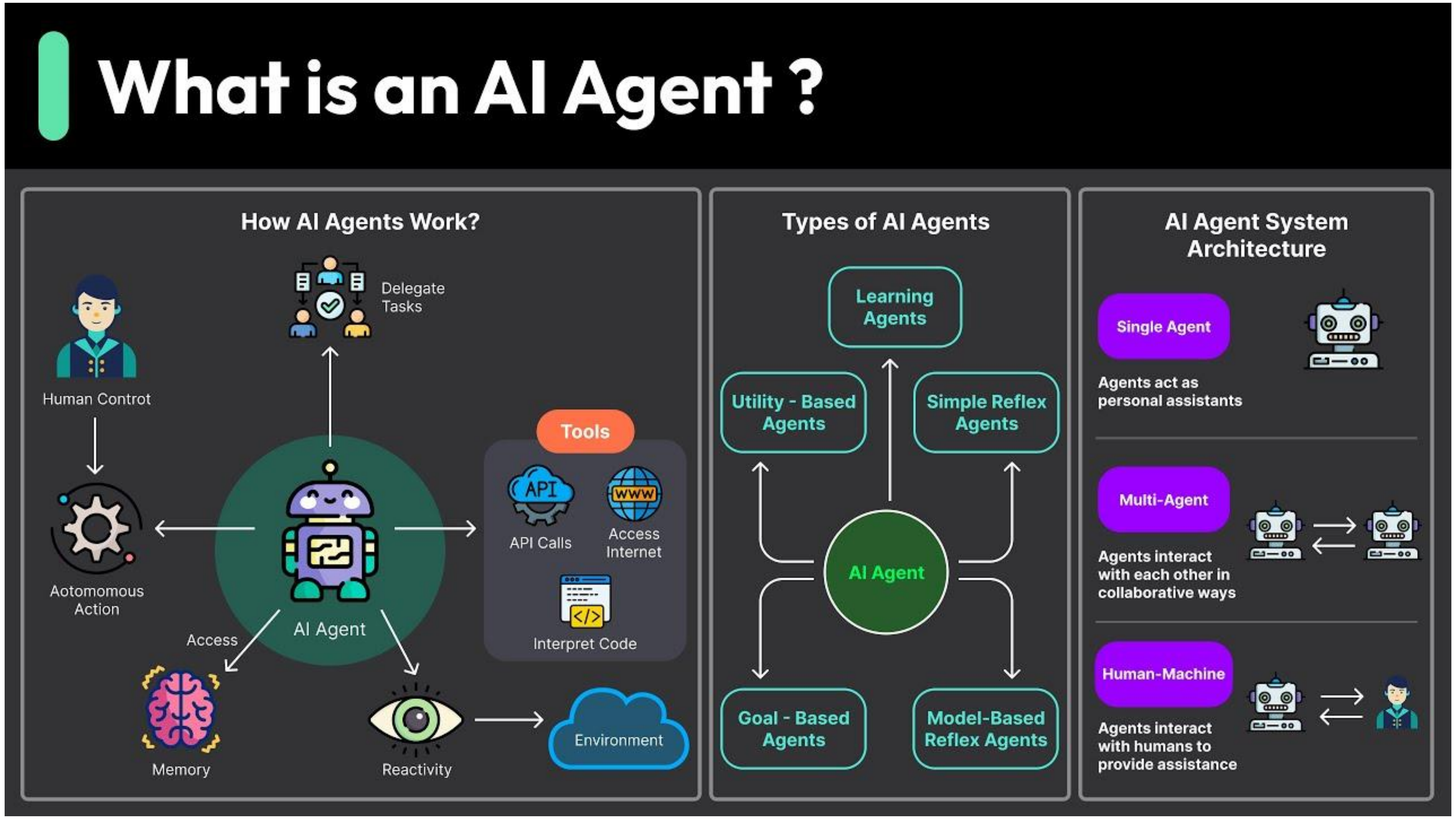
These can also be applied to coding assistants, drug discovery, create 3D environments and personalities

Generative AI also gave rise to AI agents, which can help to automate many tasks that we can do online, including creating events, purchase goods and services etc

AI agents are trending due to advances in LLMs powering them and opening the possibilities of what they can do

https://operator.chatgpt.com/

Agentic AI Agents



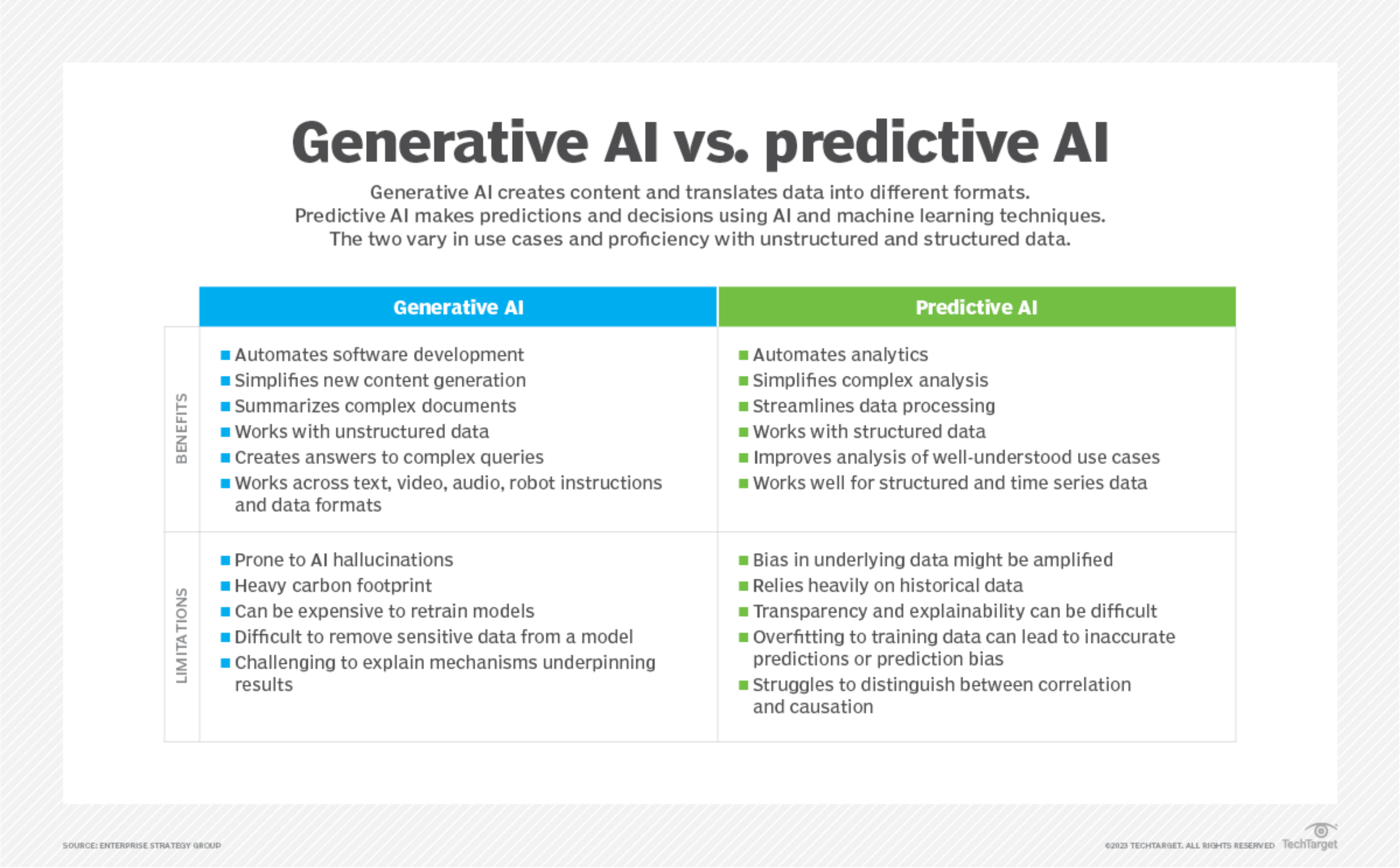
An AI agent is a system (software or robot) that perceives its environment, makes decisions, and acts to achieve specific goals

It's a smart assistant or AI worker that can autonomously perform tasks, learn from experience, and adapt to new situations

However, the potential for misuse is real as well, currently bots are being used for scams, scalping

https://www.youtube.com/watch?v=eHEHE2fpnWQ

Generative AI vs Predictive AI



Generative AI vs. predictive AI

Generative AI creates content and translates data into different formats.
Predictive AI makes predictions and decisions using AI and machine learning techniques.
The two vary in use cases and proficiency with unstructured and structured data.

	Generative AI	Predictive AI
BENEFITS	■ Automates software development ■ Simplifies new content generation ■ Summarizes complex documents ■ Works with unstructured data ■ Creates answers to complex queries ■ Works across text, video, audio, robot instructions and data formats	■ Automates analytics ■ Simplifies complex analysis ■ Streamlines data processing ■ Works with structured data ■ Improves analysis of well-understood use cases ■ Works well for structured and time series data
LIMITATIONS	■ Prone to AI hallucinations ■ Heavy carbon footprint ■ Can be expensive to retrain models ■ Difficult to remove sensitive data from a model ■ Challenging to explain mechanisms underpinning results	■ Bias in underlying data might be amplified ■ Relies heavily on historical data ■ Transparency and explainability can be difficult ■ Overfitting to training data can lead to inaccurate predictions or prediction bias ■ Struggles to distinguish between correlation and causation

SOURCE: ENTERPRISE STRATEGY GROUP

Generative AI and predictive AI vary in how they handle use cases and unstructured and structured data

Generative AI creates new data, while predictive AI makes predictions

Both have their time and place to be deployed

Explore the benefits and limitations of each, and use that to make decisions on what to use

https://www.techtarget.com/searchenterpriseai/tip/Generative-AI-vs-predictive-AI-Understanding-the-differences

When to use GenAI?

1. Content Generation at Scale
Use case: Creating text, images, or code with consistent structure
Example: Product descriptions, email replies, code autocompletion
Why GAI: High efficiency and creativity at low marginal cost

2. Unstructured Input → Structured Output
Use case: Summarizing documents, extracting meaning from messy input
Example: Summarizing customer support logs, turning audio into action points
Why GAI: Handles language and patterns better than rules-based systems

3. Tasks Needing Personalization or Variation
Use case: Personalized learning content, dynamic marketing copy
Why GAI: Can tailor outputs using embeddings, persona prompts

4. Rapid Prototyping or Ideation
Use case: Brainstorming designs, writing draft policies, generating test data
Why GAI: Speeds up human creativity cycles

When to not use GenAI?

1. When Accuracy, Reliability, or Legal Precision is Critical
Example: Tax filing, medical advice, legal contracts
Why Not: GAI can hallucinate or be non-deterministic

2. When the Output Requires Factual Integrity
Example: News reports, academic citations
Why Not: GAI may fabricate or cite non-existent sources

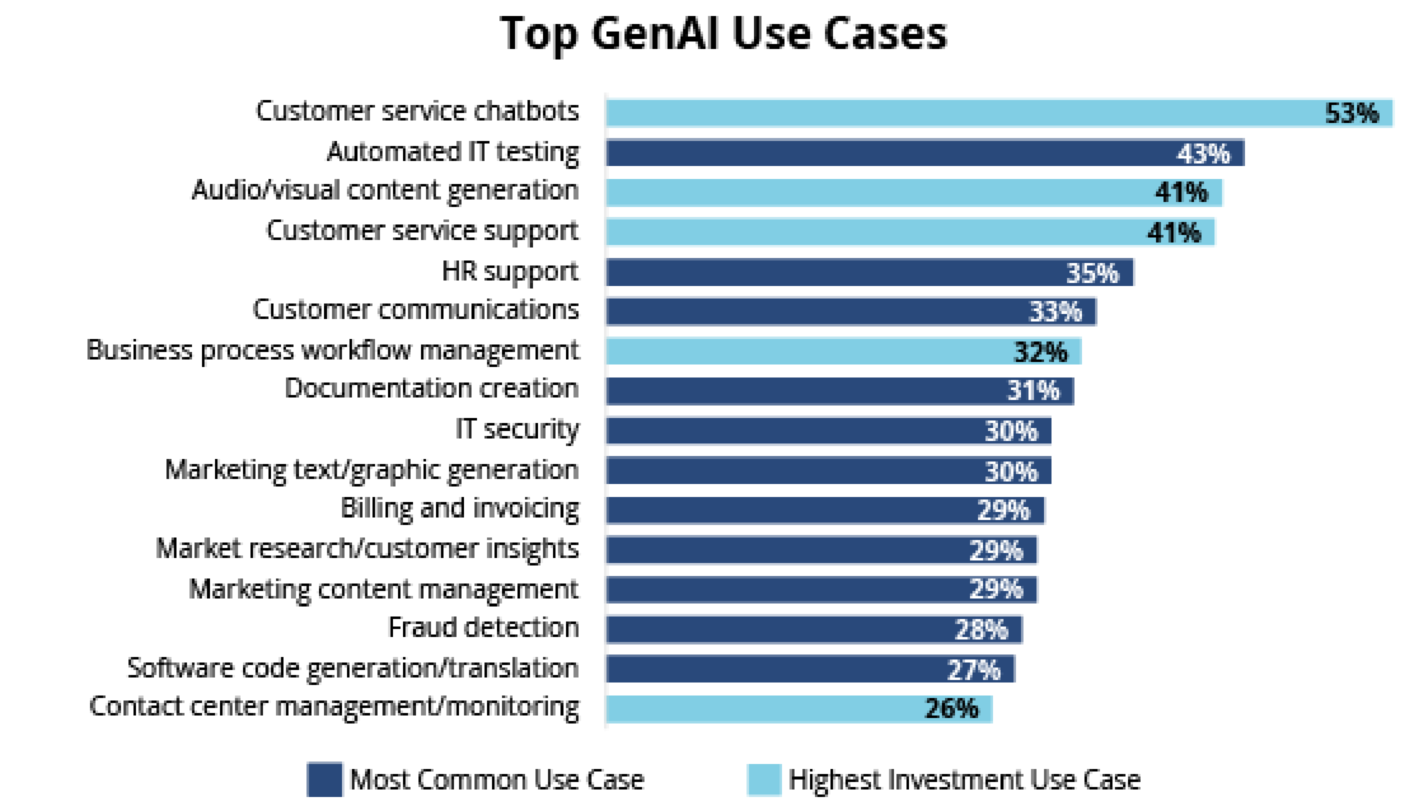
3. When You Need Traceability and Auditability
Example: Financial decision systems, safety-critical applications
Why Not: GAI is often a black box with low explainability

4. When the Task is Simple or Rule-Based
Example: Sorting emails, logging sensor data
Why Not: Simpler automation (regex, workflows) is cheaper, faster, and safer

When to not use GenAI?

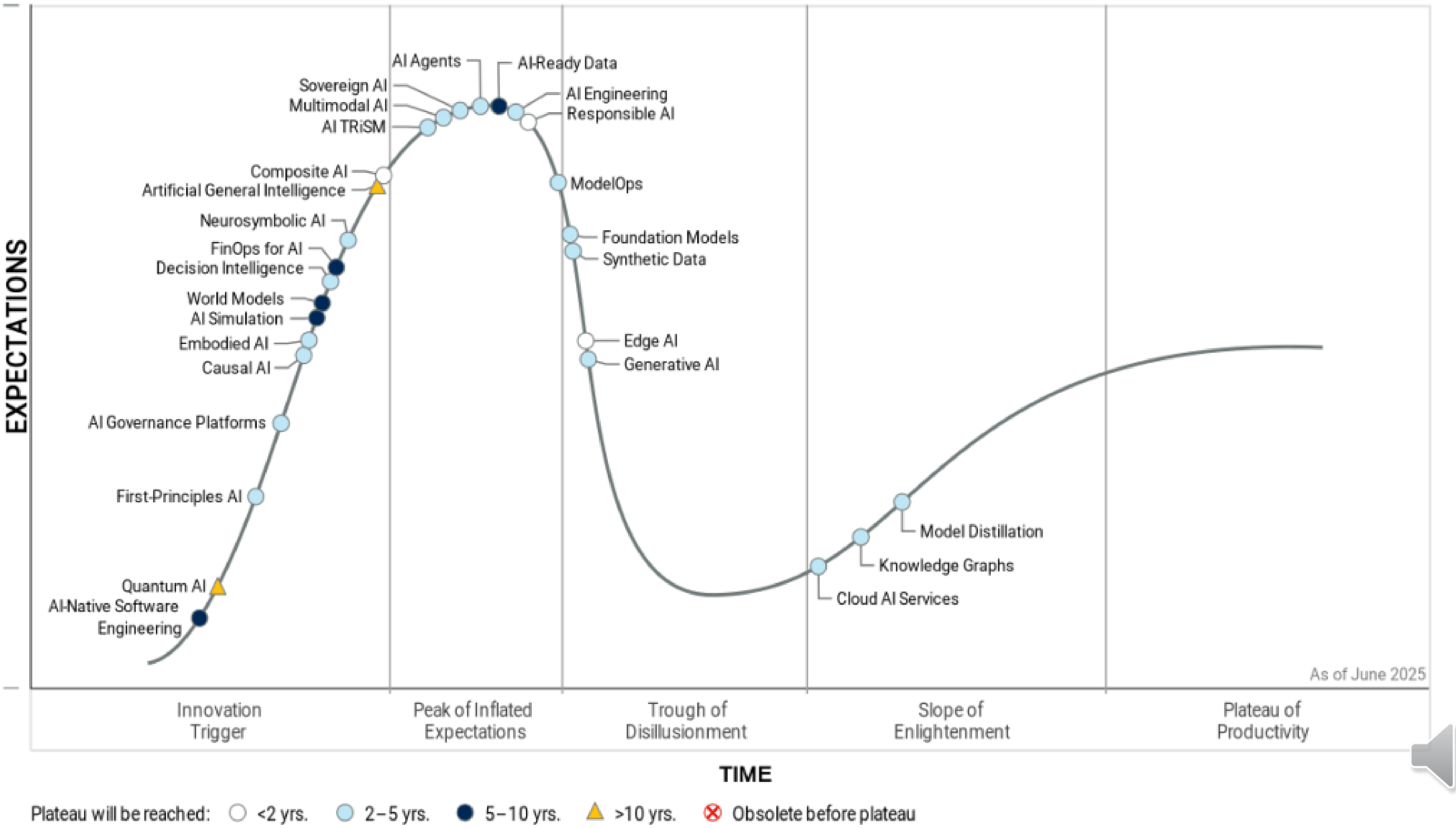
Criterion	✅ Use GAI	❌ Don’t Use GAI
Unstructured data processing	✅	
High stakes (health, legal, finance)		❌
Output needs creativity or diversity	✅	
Factual correctness is critical		❌
Needs interpretability & traceability		❌
Task is simple and repetitive		❌
You have access to high-quality data	✅	

GenAI as a Spectrum

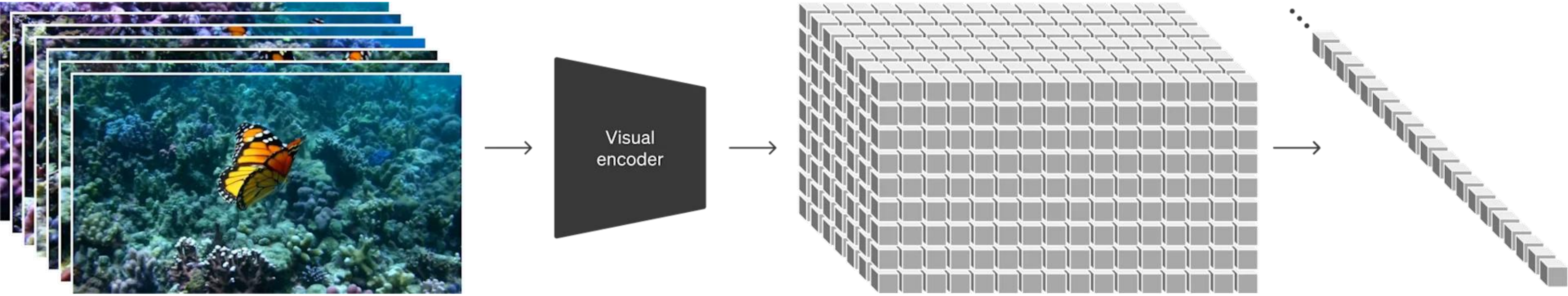


Source: ISG, 2024; Generative AI Use Case Study, n=201; Multiple Responses Allowed

Hype Cycle for Artificial Intelligence, 2025



Generative AI Use Cases: Sora



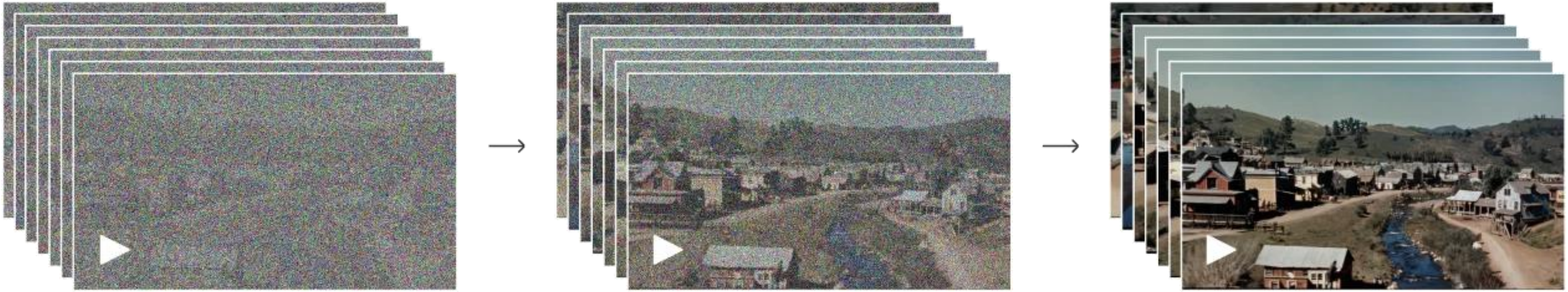
OpenAI SORA is a text to video model that generates short videos based on user prompts.

SORA uses patches to represent visual data, similar to LLMs using text tokens.

At the high level, videos are turned into patches by first compressing them into lower dimensional latent space, then decomposing into spacetime patches

https://openai.com/sora/, https://openai.com/index/video-generation-models-as-world-simulators/

Generative AI Use Cases: Sora



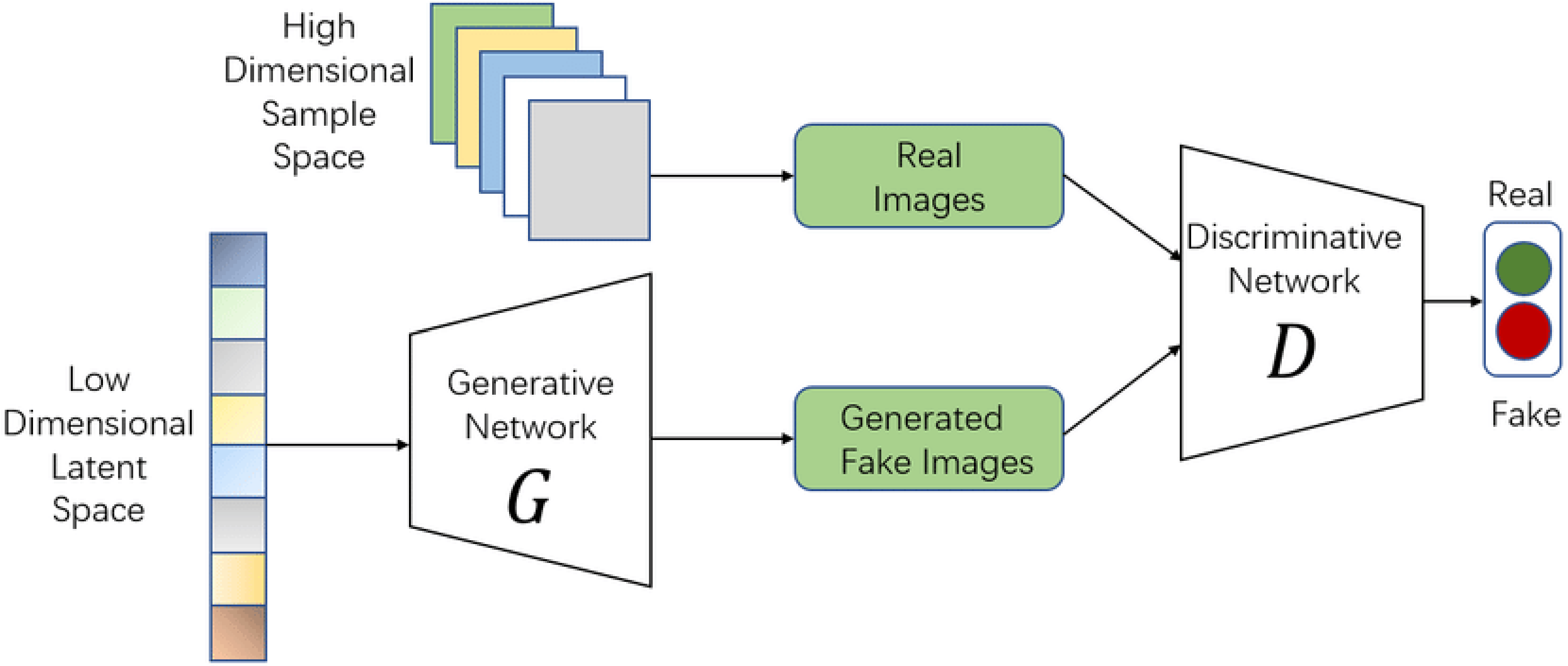
SORA is a Transformer Diffusion Model.

Diffusion Process: The model is given noisy patches and conditioning information like text prompts, then the model is trained to predict the original “clean” patches.

They are loosely inspired by the Brownian motion in physics where particles collide into each other and each step is a small random walk

https://openai.com/sora/, https://openai.com/index/video-generation-models-as-world-simulators/

Generative Adversarial Networks(GANs)

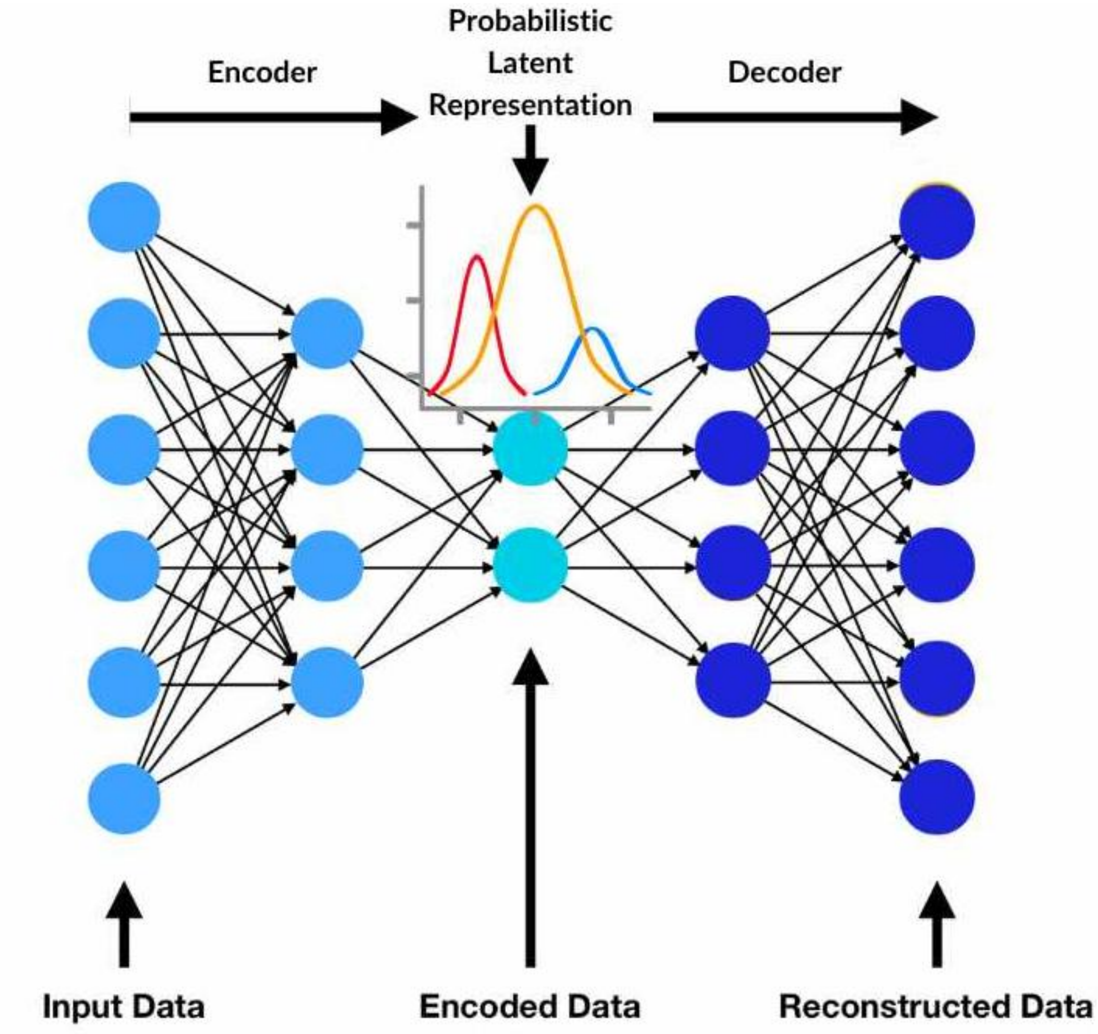


GANs consist of 2 neural networks, Discriminator and Generator that competes against each other using deep learning to scrutinize, capture and replicate variations within a dataset

GANs can create photorealistic fake images (deepfakes) that are indistinguishable from real images

https://medium.com/analytics-vidhya/a-review-of-generative-adversarial-networks-part-1-a3e5757a3dc2

Variational Autoencoders(VAEs)



VAEs are a type of neural network designed to learn efficient data representations

This allows for dimensionality reduction or feature learning, and subsequently new similar data

Encoder maps the input data to a latent space

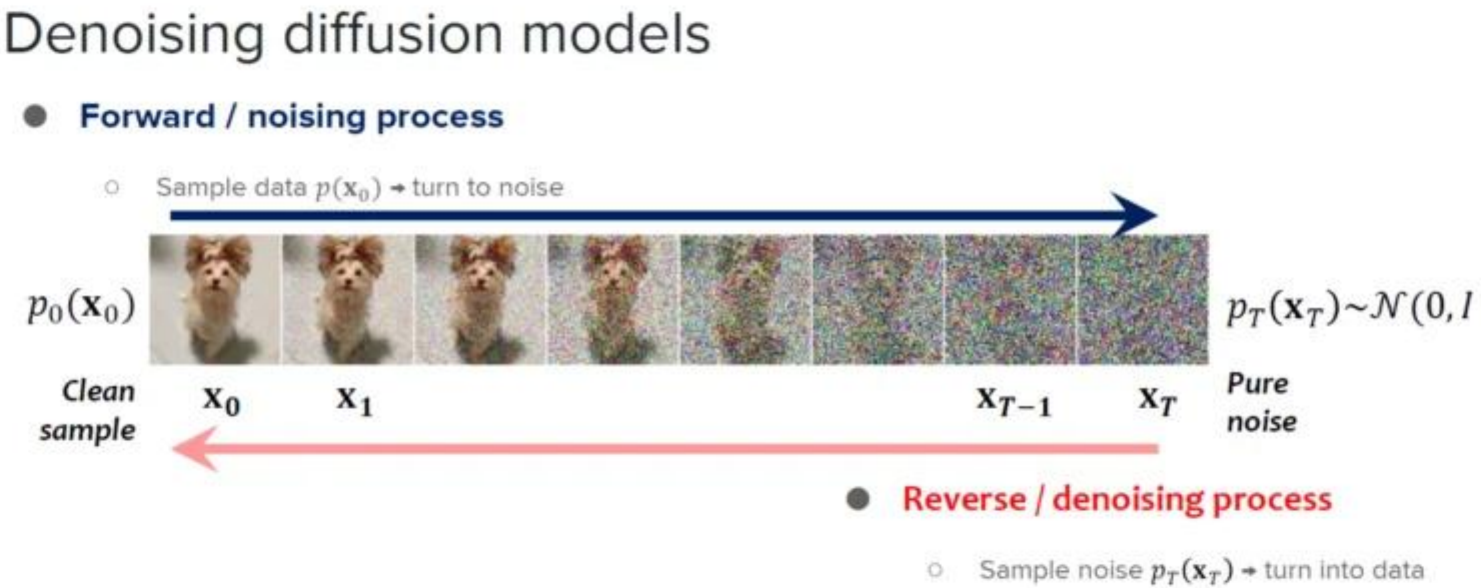
Latent space is represents the compressed input data in a probabilistic manner

Decoder reconstructs data from latent space and introduces variations

The primary goal is to minimize the difference between input and reconstructed output, while incorporating randomness by sampling from latent space

https://www.datacamp.com/tutorial/variational-autoencoders

Diffusion Models



Denoising diffusion models

- Forward / noising process
  - Sample data $p(\mathbf{x}_0)$ → turn to noise
- Reverse / denoising process
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

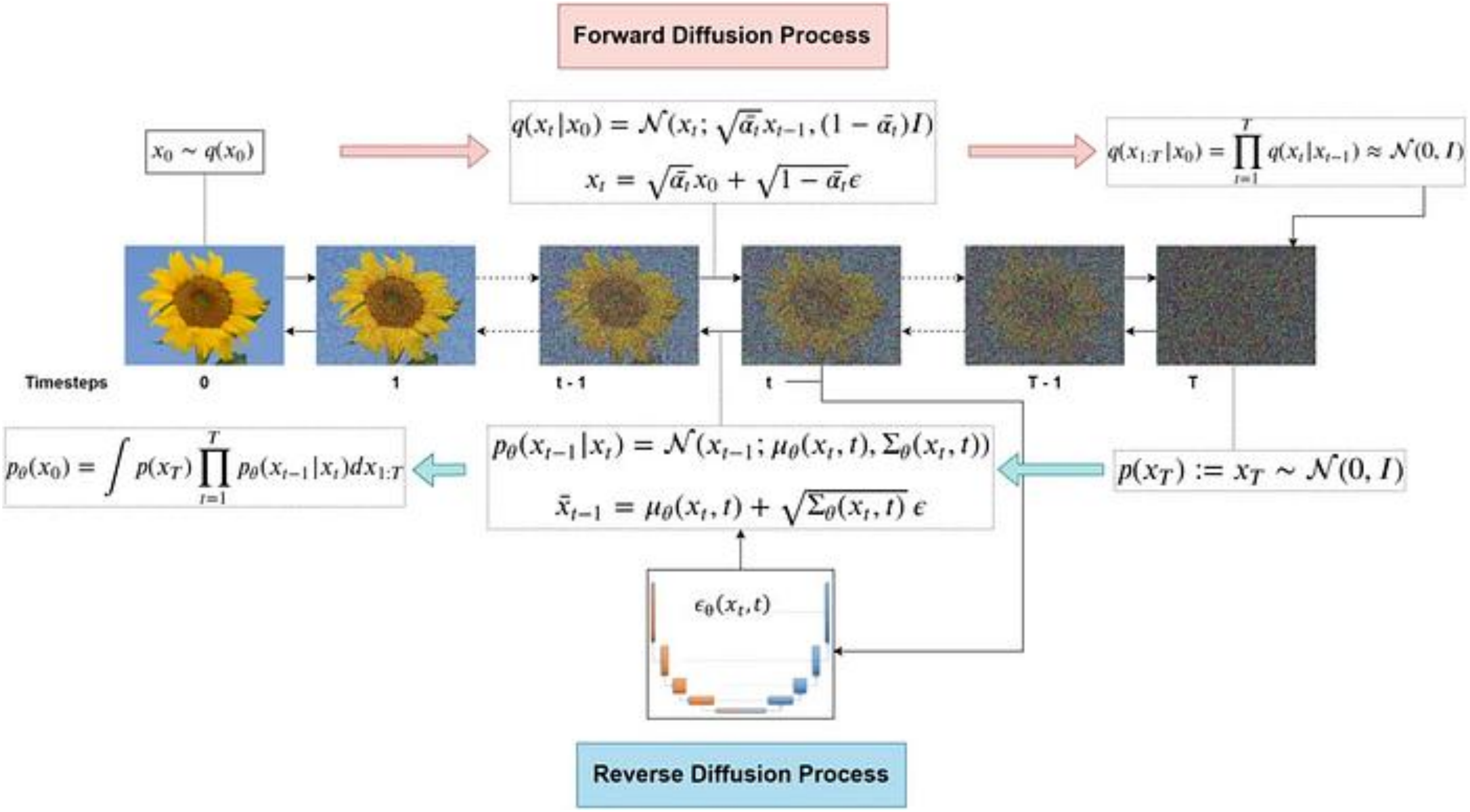
Diffusion models start with a dataset and add noise to it incrementally, until it is indistinguishable from actual noise

Reverse diffusion then takes the noisy data and progressively reconstructs the data step by step

After many iterations, the results generated by diffusion models can become high quality, photorealistic pictures

https://colab.research.google.com/drive/1aSQTgoqmyqGpLI9q7IRDIXXeMdAG-E4X?usp=sharing#scrollTo=BxG5BU-wLokJ

Diffusion Transformers


Forward Diffusion Process

$x_0 \sim q(x_0)$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_{t-1}, (1-\bar{\alpha}_t)I)$$
$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \approx \mathcal{N}(0, I)$$

$$p_\theta(x_0) = \int p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)dx_{1:T}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$
$$\bar{x}_{t-1} = \mu_\theta(x_t, t) + \sqrt{\Sigma_\theta(x_t, t)}\,\epsilon$$

$$p(x_T) := x_T \sim \mathcal{N}(0, I)$$

$\epsilon_\theta(x_t,t)$

Reverse Diffusion Process

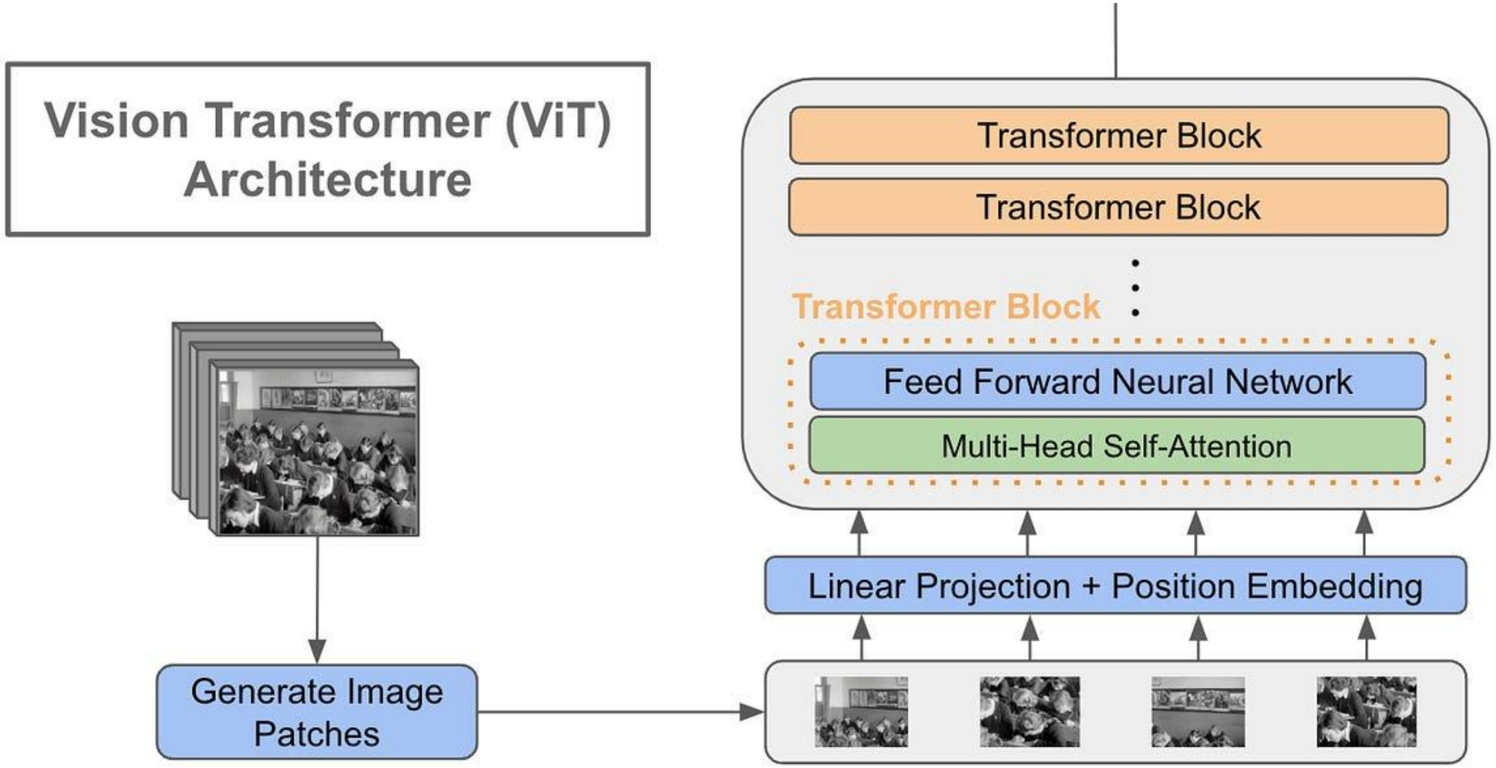
Diffusion Transformers generate data by adding noise gradually to a target image, then reversing that process guided by a transformer.

The transformer guides the process by:

1. Modelling noise, it processes the noisy data as a sequence
2. Self-attention to look at different parts of the input and find signal vs noise
3. Position and timestamp
4. Trained to predict the noise at every time step

https://medium.com/@threehappyer/understanding-dit-diffusion-transformer-in-one-article-2f7c330ad0ea

Vision Transformers (Classification)



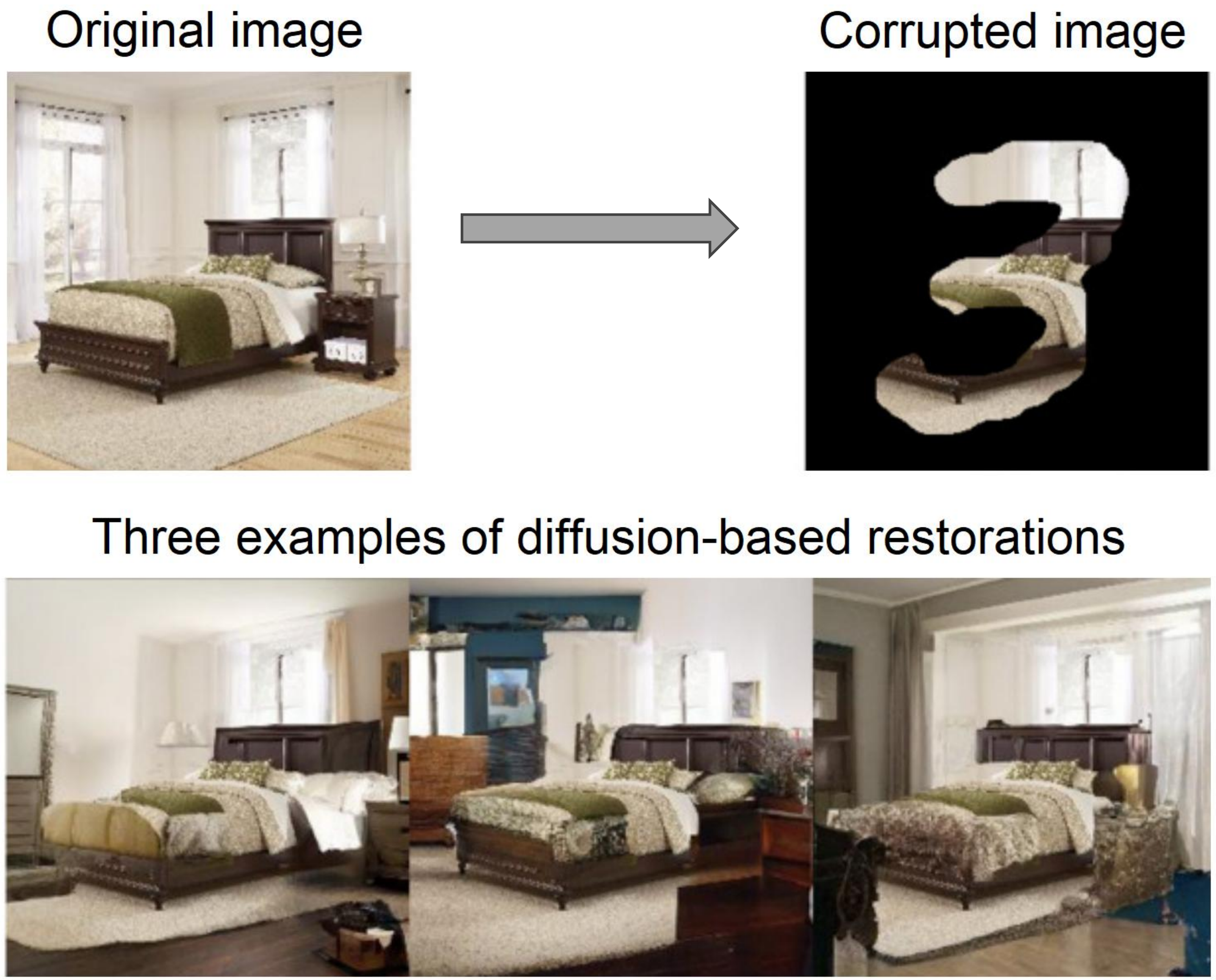
Vision Transformer (ViT) Architecture
Transformer Block
Transformer Block
Transformer Block
Feed Forward Neural Network
Multi-Head Self-Attention
Linear Projection + Position Embedding
Generate Image Patches

Vision Transformers work by splitting an image into patches, treats them like tokens instead of using convolutions

1. Image is flattened from a matrix into a vector (like word embeddings in NLP)
2. Positional encodings are added to retain position of each patch
3. A classification token (CLS) is prepended to sequence, output embedding of CLS use for final classification
4. Embedings are passed through self attention layers to learn relationships
5. Finally CLS output class label

https://medium.com/@threehappyer/understanding-dit-diffusion-transformer-in-one-article-2f7c330ad0ea

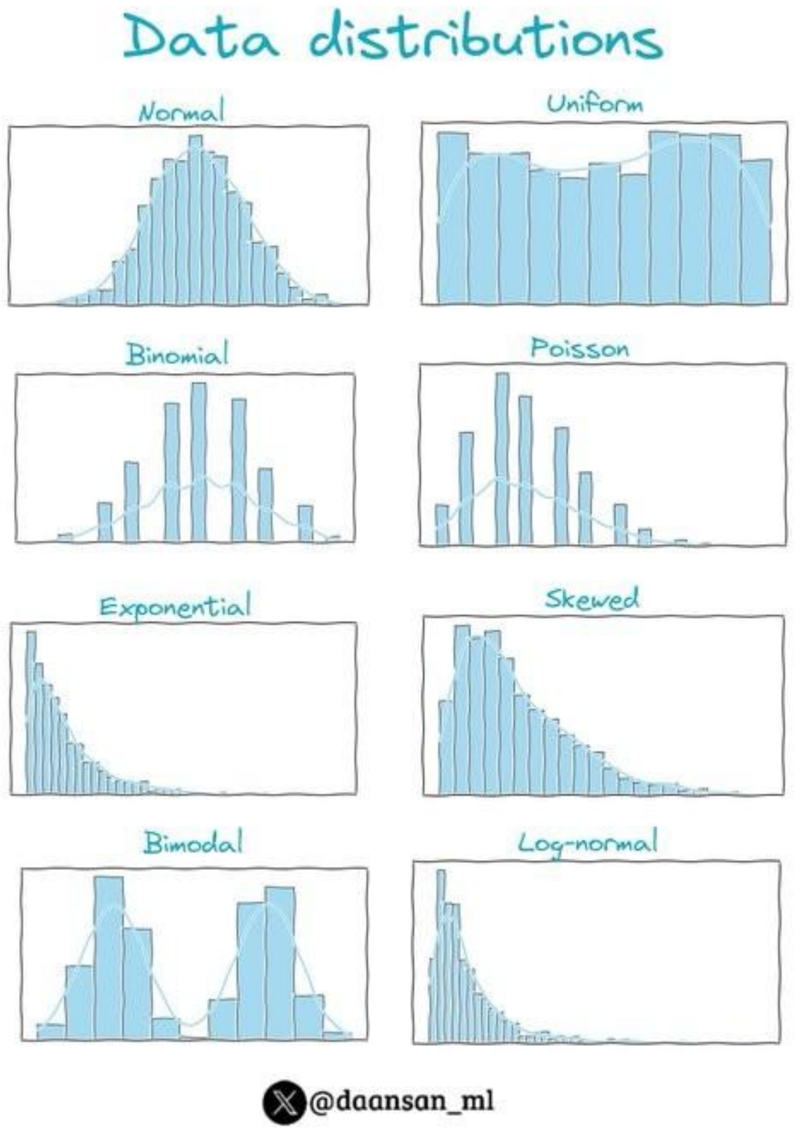
Diffusion Models



Diffusion models can sometimes surprise us with their creativity

https://colab.research.google.com/drive/1aSQTgoqmyqGpLI9q7IRDIXXeMdAG-E4X?usp=sharing#scrollTo=BxG5BU-wLokJ

Data Distribution



A data distribution is a graphical representation of data, showing how they are spread across a range of values and how they vary.

It helps us understand the characteristics of data, such as mean, spread, modality (no of peaks) and shape

We use data visualisation to visually show data distribution

Gen AI models uses data distribution to learn these characteristics of the data and create similar synthetic data

It is similar to randomly sampling from a distribution, where each data point is a sample

How can GenAI help you?

AI can help you to learn better by doing the following to adapt to you

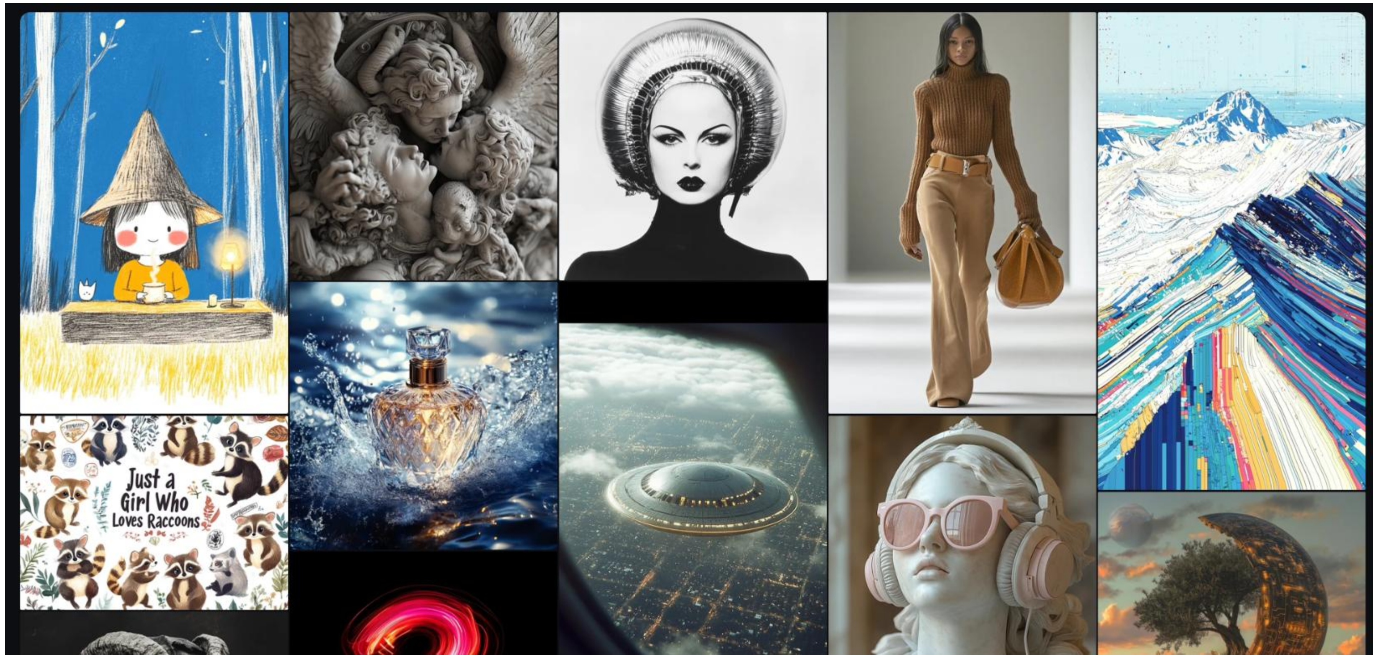
1. Generating materials tailored to your strengths and preferences, avoiding your weaknesses
2. Providing simple terms for beginners, or advanced details for experts
3. Creating interactive learning tools such as simulations, role playing to make learning fun
4. You can use ChatGPT to test if your understanding is correct and provide alternative perspectives
5. GenAI can help you to draft content by creating a outline or ideas

Example:

Use ChatGPT to find out what are the core ideas in Deep Learning, Computer Vision and NLP, and learn those concepts thoroughly before proceeding to do hands on

Bottomline: embrace and use AI to your advantage, and you will be better off compared to others who do not understand AI

Generative AI: Ethical Considerations



GenAI can now imitate human creativity, create content and show remarkable capabilities.

It can help us to discover new drugs, enhance understanding and predictions about climate change,

GenAI can design optimized hardware and software systems, creating a feedback loop for better AI

… However, we need Ethics, Safety and Governance systems to regulate AI

https://www.midjourney.com/explore?tab=top

Generative AI: Ethical Considerations

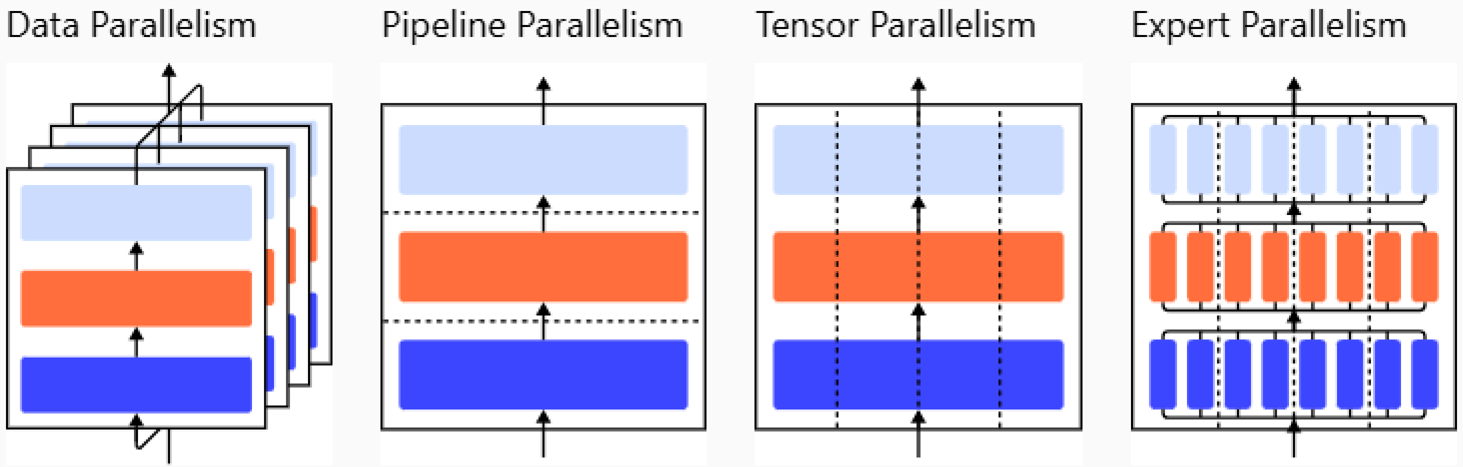
Bias Mitigation
Generative AI models are trained on large datasets that may contain biases, leading to outputs that could perpetuate stereotypes or discriminate against certain groups.
Efforts will focus on reducing biases in AI outputs and ensuring fairness across diverse demographics.

Regulation and Safety of Use
Generative AI can be misused for harmful purposes like creating malicious software, phishing emails, or harmful propaganda.
Regulation is key to minimize Governments and organizations are likely to develop and enforce standards to ensure responsible use.

Privacy and Data Security
Generative AI systems often rely on user data to improve personalization, raising concerns about data privacy and security.
Measures such as removing personally identifiable info, secure storage and encryption is needed for privacy

https://openai.com/index/techniques-for-training-large-neural-networks/

Training Large Models



An illustration of various parallelism strategies on a three-layer model. Each color refers to one layer and dashed lines separate different GPUs.

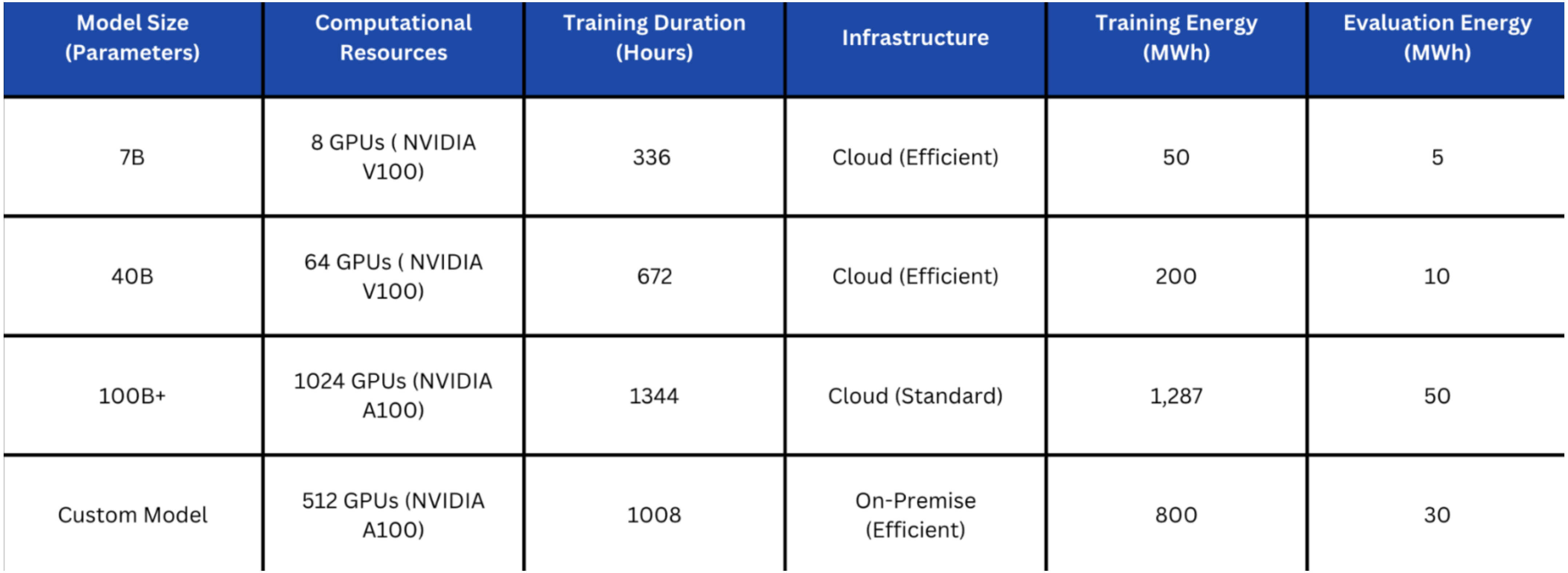
Training large models is difficult, both engineering and research challenge which requires orchestrating a cluster of GPUs to collectively train together

This process of collectively training together is known as parallelism, key to accelerating the process

Training is both a data and software engineering problem, where AI engineers find many ways to optimise the process

https://openai.com/index/techniques-for-training-large-neural-networks/

Energy Footprint of LLMs



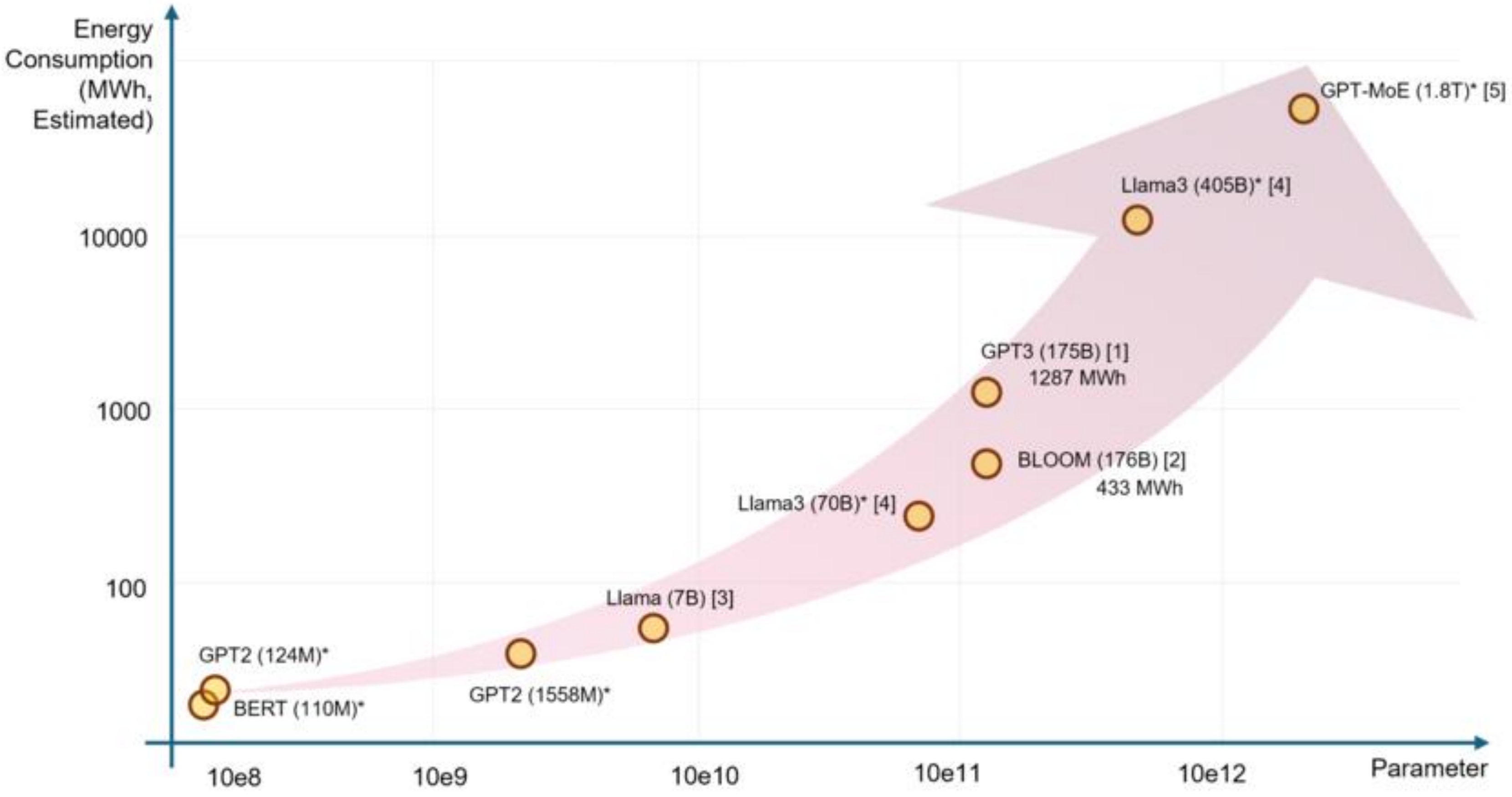
Model Size (Parameters)	Computational Resources	Training Duration (Hours)	Infrastructure	Training Energy (MWh)	Evaluation Energy (MWh)
7B	8 GPUs (NVIDIA V100)	336	Cloud (Efficient)	50	5
40B	64 GPUs (NVIDIA V100)	672	Cloud (Efficient)	200	10
100B+	1024 GPUs (NVIDIA A100)	1344	Cloud (Standard)	1,287	50
Custom Model	512 GPUs (NVIDIA A100)	1008	On-Premise (Efficient)	800	30

Training these massive LLMs requires extensive computation resources often involve thousands of GPUs running for weeks/months

The smallest 7B parameter LLM requires at least 55 MWh of energy, excluding energy consumption during deployment and user prompts

https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/

Energy Footprint of LLMs



A prompt on ChatGPT costs nearly 10 times more than a search on Google.

As the energy consumption of AI grows exponentially, we are looking at alternative sources of energy to power them

One area that big tech is exploring at is nuclear power:

- Consistency and Reliability
- Low carbon footprint
- Scalable

https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/

References for Chapter 6

https://openai.com/sora/

https://openai.com/index/video-generation-models-as-world-simulators/

https://medium.com/analytics-vidhya/a-review-of-generative-adversarial-networks-part-1-a3e5757a3dc2

https://www.datacamp.com/tutorial/variational-autoencoders

https://colab.research.google.com/drive/1aSQTgoqmyqGpLI9q7IRDlXXeMdAG-E4X?usp=sharing#scrollTo=BxG5BU-wLokJ

https://deeprevision.github.io/posts/001-transformer/

https://openai.com/index/techniques-for-training-large-neural-networks/

https://www.midjourney.com/explore?tab=top

https://openai.com/index/techniques-for-training-large-neural-networks/

Note: All online articles were accessed between Oct to Dec 2024

The End
Questions?