Cornellius Yudha Wijaya

40 POPULAR Generative AI Interview Questions and Answers?

NON BRAND DATA

What is Generative AI?

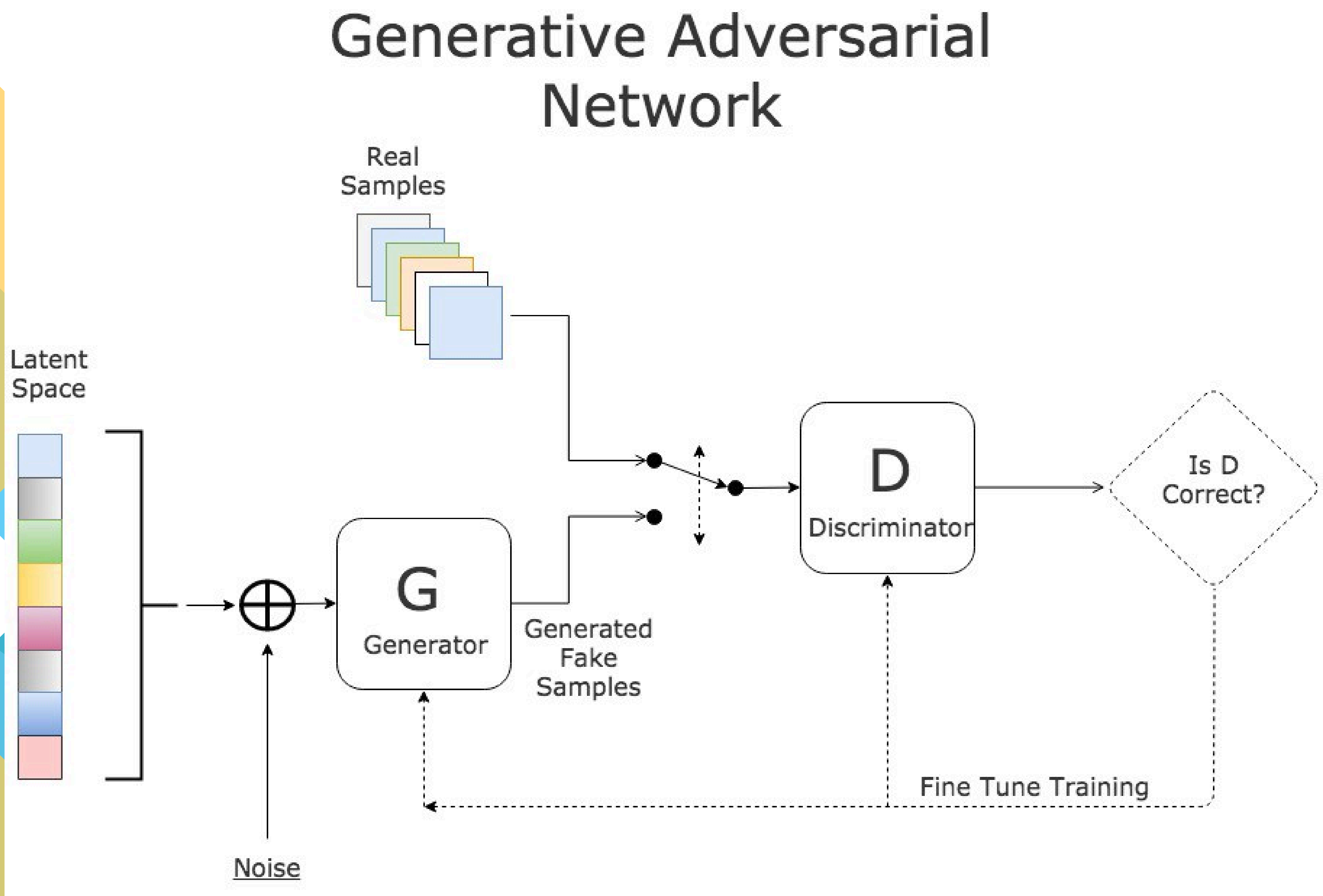
Generative AI refers to artificial intelligence systems designed to create content, such as images, text, audio, or video, by learning patterns from existing data and generating new, original content.

How is Generative AI different from discriminative AI?

- Generative models learn joint probability distributions and create new data instances.
- Discriminative models classify data into different categories based on features and labels.

What are GANs? Explain their architecture.

Generative Adversarial Networks consist of two neural networks: a Generator (creates synthetic data) and a Discriminator (distinguishes real data from fake). Both networks compete against each other, improving the quality of generated data.

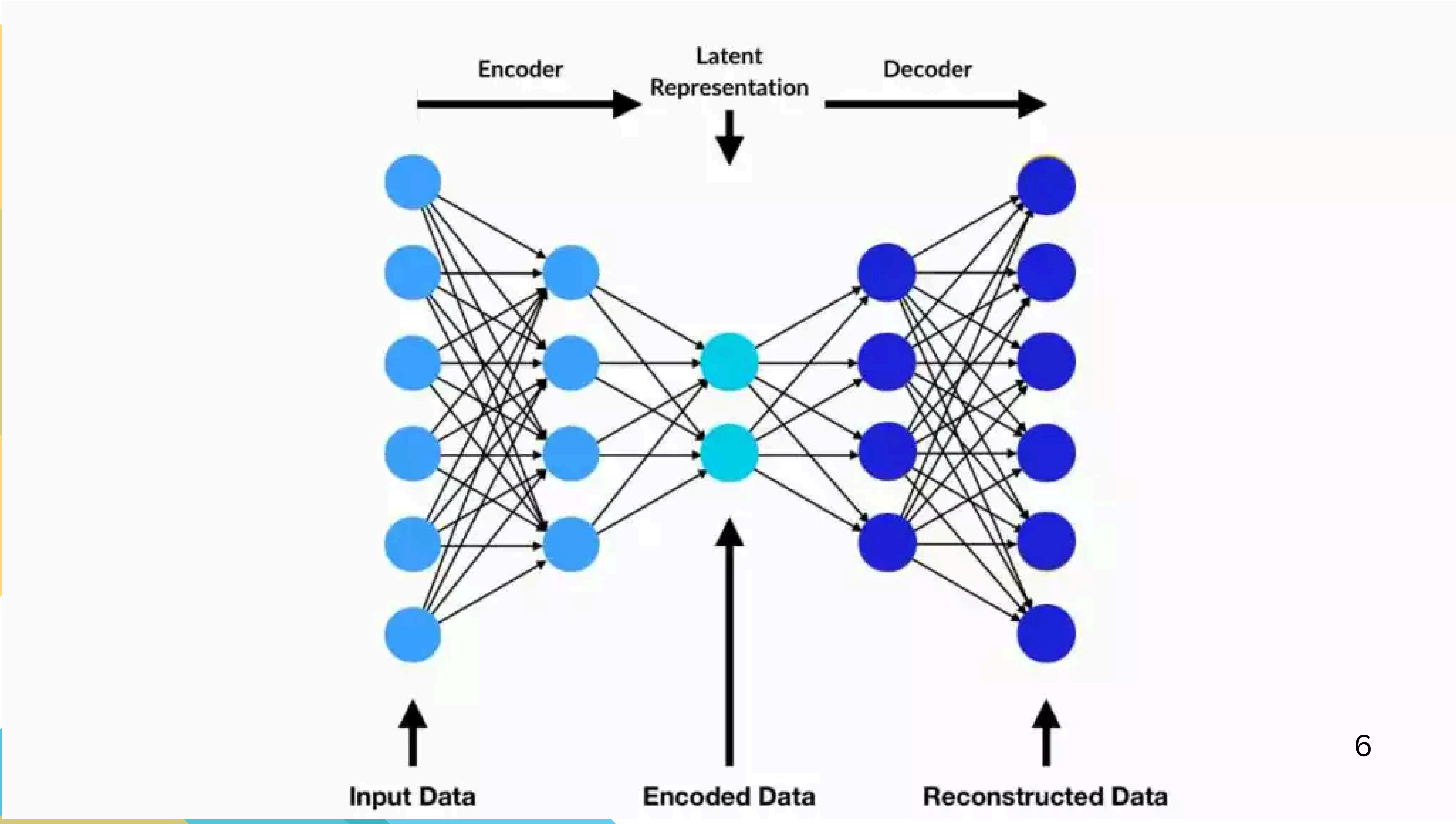


What is a conditional GAN (cGAN)?

Conditional GANs generate data based on specified conditions or labels, allowing targeted data generation. It is an enhanced version of GANs where both the generator and discriminator are conditioned on some additional information.

What is Variational Autoencoders?

VAEs are generative models using an encoder-decoder architecture. They encode inputs into a latent space, enabling generation of new, diverse outputs through latent space sampling.



Difference between GANs and VAEs?

- GANs produce sharper, realistic outputs but are harder to train.
- VAEs are easier to train but produce smoother outputs with less sharpness.

What is a Diffusion Model?

Diffusion models generate data by reversing a diffusion (noise-adding) process, gradually transforming random noise into structured output.

What are Transformer models, and why are they important for generative AI?

Transformers use attention mechanisms to model sequences. They are effective in generative AI tasks like language generation due to capturing long-range dependencies.

How does GPT generate human-like text?

GPT uses next-token prediction trained on vast datasets, leveraging attention mechanisms to capture context and generate coherent text.

What are LLMs, and how are they trained?

LLMs are large-scale neural networks trained on extensive textual data, using self-supervised techniques such as next-word prediction and masked-language modeling.

What is fine-tuning in LLMs?

Fine-tuning adapts a pretrained LLM to specific tasks or datasets, improving performance for targeted use cases.

What is Prompt Engineering?

Prompt engineering is crafting effective prompts to guide generative AI models to produce desired responses or outputs.

What is "temperature" in language models?

Temperature controls randomness in generation. Lower temperatures (e.g., 0.2) yield deterministic outputs; higher values (e.g., 1.0) produce diverse and creative outputs.

What is top-k sampling?

Top-k sampling randomly selects the next token from the top K most probable tokens to maintain diversity and coherence.

How does reinforcement learning from human feedback (RLHF) improve generative models?

RLHF aligns AI outputs with human preferences by fine-tuning models using human-ranked outputs as rewards.

What's zero-shot vs few-shot prompting?

- Zero-shot: Tasks performed without explicit examples.
- Few-shot: Providing a few examples within the prompt.

What are latent spaces?

Latent spaces are compressed representations of data points, capturing key features for efficient data manipulation and generation.

What is multimodal generative AI?

Multimodal generative AI creates outputs across different modalities (e.g., text-to-image, text-to-audio).

What's tokenization in LLMs?

Breaking input text into discrete tokens (words or subwords) for processing by language models.

What's tokenization in LLMs?

Breaking input text into discrete tokens (words or subwords) for processing by language models.

What is token embedding?

Token embedding refers to the process of converting discrete text tokens (e.g., words, subwords, characters) into continuous numerical vectors.

What are Neural Radiance Fields (NeRFs)?

Neural Radiance Fields (NeRFs) are generative models capable of synthesizing highly realistic and novel views of complex 3D scenes based purely on a set of 2D images taken from different angles.

Why are diffusion models popular in image generation?

Diffusion models have become popular in image generation due to their remarkable ability to generate high-quality, realistic, and diverse images. Unlike GANs, which train two competing networks simultaneously and often suffer from training instability (e.g., mode collapse), diffusion models offer a stable training process.

What is the attention mechanism?

The attention mechanism is a neural network component enabling models to selectively focus on specific parts of the input data when generating outputs. Instead of processing inputs uniformly, attention assigns weights to inputs, emphasizing important or relevant portions dynamically.

What is the DreamBooth technique?

DreamBooth is a fine-tuning method enabling personalized generation using diffusion models by adapting pretrained generative models to specific subjects or objects using very few images (often 3–5 images).

What are adversarial examples?

Adversarial examples are intentionally crafted inputs designed to deceive machine learning models, causing incorrect predictions or outputs. Typically, these inputs appear indistinguishable from genuine data to human observers but exploit subtle vulnerabilities in the model's learned representations.

What is the KL divergence?

Kullback-Leibler (KL) divergence is a statistical measure quantifying the difference or distance between two probability distributions. In generative AI models, KL divergence is commonly used as a regularization term. Minimizing KL divergence ensures generated samples align closely with target distributions, leading to more realistic outputs.

What is Teacher Forcing in text generation?

Teacher forcing trains generative models by conditioning predictions on correct prior tokens during training. It speeds convergence but can create discrepancies (exposure bias) between training and inference phases.

What is Beam Search in text generation?

Beam search generates text by maintaining multiple candidate sequences simultaneously. It explores several most promising candidate sequences (beams) at each step, balancing diversity and computational efficiency, typically yielding more coherent outputs than greedy search.

How to mitigate hallucinations in LLMs?

- Prompt engineering (clear instructions)
- Use retrieval-augmented generation (RAG)
- Fact-checking modules and validation layers
- Fine-tuning with high-quality, verified datasets

What is a Hypernetwork?

A Hypernetwork is a neural network generating weights for another neural network, allowing quick adaptation of models to different tasks or data distributions without extensive retraining.

What is a Hypernetwork?

A Hypernetwork is a neural network generating weights for another neural network, allowing quick adaptation of models to different tasks or data distributions without extensive retraining. Useful for rapid fine-tuning, personalization, or adaptive learning scenarios.

What is a Hypernetwork?

A Hypernetwork is a neural network generating weights for another neural network, allowing quick adaptation of models to different tasks or data distributions without extensive retraining. Useful for rapid fine-tuning, personalization, or adaptive learning scenarios.

How does Contrastive Learning Work?

Contrastive learning trains models by contrasting similar (positive) and dissimilar (negative) data examples. It encourages the neural network to learn embeddings where related inputs (e.g., an image and its caption) become close in the latent embedding space, while unrelated pairs become distant.

What Are Challenges in Real-Time Generative AI?

- Latency Constraints.
- High Computational Requirements
- Inference Efficiency:

What Are Challenges in Real-Time Generative AI?

- Latency Constraints.
- High Computational Requirements
- Inference Efficiency:

How do you handle safety concerns in deployment?

- Content filters
- Human moderation
- Continuous monitoring and updates

What are the Popular Libraries for Generative AI?

- PyTorch
- TensorFlow
- Hugging Face Transformers
- OpenAI API
- Diffusers.

How do you ensure reproducibility in generative AI?

- Fix random seeds
- Version control models and data
- Documenting hyperparameters and environment

What future directions do you see in generative AI?

- Enhanced multimodality
- Alignment improvements
- Safer deployment
- Increased model efficiency
- Greater personalization.