A to Z

Must know
Gen AI Terms

A

1. **Agents:** Software robots that can independently perceive and act within their environment to achieve goals, like booking flights or navigating virtual worlds. Imagine a tiny AI assistant helping you manage your online life.
2. **AGI (Artificial General Intelligence):** A hypothetical AI capable of understanding and learning any intellectual task a human can, achieving human-level performance across various domains, not just one specific task. Think of a super-intelligent machine that can write poetry, diagnose diseases, and compose symphonies.
3. **Alignment:** Ensuring AI goals and values are compatible with human values, preventing harmful or unintended consequences. It's like training a puppy to understand what behavior is good and bad for the household.
4. **Attention:** Mechanisms in neural networks that selectively focus on important parts of the input data, similar to how you might pay attention to a specific speaker during a crowded conversation.
5. **Autoencoders:** Neural networks that learn compressed representations of data and then reconstruct the original data from those representations, like a secret code for images or music.

B

6. **Back Propagation:** An algorithm that lets neural networks learn by figuring out how much to adjust their internal connections based on how well they perform on a task, like a student correcting their mistakes based on feedback.
7. **Bias:** Assumptions baked into AI models, often unintentionally, that can lead to unfair or discriminatory outcomes. It's like a faulty scale always tipping slightly to one side, skewing the results.

8. **BigGAN:** A powerful type of GAN known for generating incredibly realistic and high-resolution images, like painting you a picture so real it feels like you can step into it.

C

9. **Capsule Networks:** Networks using capsules instead of neurons to capture spatial relationships and parts of objects, like recognizing a chair with legs and back, not just a blob of pixels.
10. **Chain of Thought:** A proposed way for AI models to explain their reasoning process by showing a sequence of intermediate conclusions leading to the final decision, like tracing the steps in a mathematical proof.
11. **Chatbot:** Computer programs designed to simulate conversation with humans, like your friendly virtual assistant answering your questions and booking appointments.
12. **ChatGPT:** OpenAI’s large language model known for its ability to generate human-quality text and engage in open-ended conversations, like chatting with a witty and knowledgeable friend.
13. **CLIP (Contrastive Language–Image Pretraining):** An AI model that can connect text and images, understanding what a picture is about and describing it accurately, like a multilingual translator for visual language.
14. **CNN (Convolutional Neural Network):** Models specialized in processing data arranged in grids, like images, by identifying patterns and features within them, similar to how your eyes scan a picture to recognize objects.
15. **Conditional GAN (cGAN):** A GAN that can generate data based on specific additional information, like creating faces that fit a certain age or mood, like a fashion designer with unlimited fabric and imagination.

16. **CycleGAN:** A model that translates images from one style to another without needing paired examples, like transforming a cityscape into a watercolor painting.

D

17. **Data Augmentation:** Artificially increasing the amount and diversity of training data to make AI models more robust and generalizable, like giving a student a variety of practice problems to prepare for a test.
18. **DeepSpeed:** DeepSpeed is a system for training large language models on distributed systems. It uses a variety of techniques to improve the efficiency and scalability of training, such as ZeRO-Offloading, Megatron-Turing NLG, and mixed precision training. DeepSpeed has been shown to significantly reduce the time and cost of training large language models.
19. **Diffusion Models:** A new technique for generating data by gradually adding and then reversing noise, like slowly revealing a hidden image by erasing random brushstrokes.
20. **Double Descent:** A phenomenon where increasing the complexity of an AI model can initially hurt its performance before eventually improving it, like a rollercoaster with dips and rises before the final climb.

E

21. **Emergence/Emergent Behavior:** Complex and unexpected behavior arising from the interaction of simple rules in an AI system, like ants forming intricate patterns while following individual instructions.
22. **Expert Systems:** AI applications built with deep knowledge of a specific domain, like a medical diagnosis system drawing on vast medical databases.

F

23. **Few-Shot Learning:** This method trains models on a very small amount of data, typically a few examples per class. It’s designed to quickly adapt to new tasks with limited information, balancing the need for accuracy with the challenge of minimal data.
24. **Fine-tuning:** Adapting a pre-trained AI model to a specific task by further training it on smaller amounts of relevant data, like customizing a general tool to tackle a particular job.
25. **Forward Propagation:** The process in neural networks where input data flows through the network layers, transforming and generating the final output, like a recipe where ingredients go through different steps to create a dish.
26. **Foundation Model:** A large and adaptable AI model serving as a base for developing various specialized applications, like a versatile building block for different kinds of tools.

G

27. **GAN (General Adversarial Network):** A type of AI where two models compete, one generating data and the other trying to distinguish it from real data, leading to increasingly realistic and sophisticated outputs, like two artists pushing each other to create better work.
28. **Generative AI:** Machine learning models capable of autonomously creating new content, such as images, text, music, or code. Unlike traditional AI models that analyze or classify data, generative AI focuses on creative exploration and output.
29. **GPT (Generative Pretrained Transformer):** A large language model developed by OpenAI, known for its ability to generate human-quality text, translate languages, and write different kinds of creative content. It is trained on a massive dataset of text and code, allowing it to learn complex patterns and relationships within language.

30. **GPU (Graphics Processing Unit):** Specialized microprocessors designed for parallel processing, making them ideal for handling the computationally intensive tasks involved in AI training and inference. GPUs excel at computations involving arrays of data, which are common in image and video processing, data mining, and scientific computing.
31. **Gradient Descent:** An optimization algorithm used to improve the performance of machine learning models. It works by iteratively adjusting the model's internal parameters in the direction that minimizes a loss function, which measures the difference between the model's predictions and the true values.

H

32. **Hallucination/Hallucination:** When AI models generate unrealistic or nonsensical content due to limitations in their training data, biases, or incomplete understanding of the task. This can include generating images of objects that don't exist, writing text that doesn't make sense, or making predictions that are factually incorrect.
33. **Hidden Layer:** Layers in neural networks that are not directly connected to the input or output. These layers perform complex transformations on the data by learning internal representations that capture hidden patterns and relationships within the data. The number and structure of hidden layers play a crucial role in the capabilities and performance of neural networks.
34. **Hyperparameter Tuning:** Adjusting settings in a machine learning model, such as the learning rate, number of hidden layers, or regularization parameters, to achieve optimal performance. Tuning these hyperparameters is crucial for finding the right balance between model complexity and generalizability.

I

35. **Instruction Tuning:** Fine-tuning a pre-trained machine learning model by further training it on a smaller dataset that includes specific instructions or guidelines. This can be used to adapt the model to a new task or improve its performance on a specific aspect of the original task.

L

36. **Large Language Model (LLM):** A machine learning model trained on a massive dataset of text and code, capable of generating human-quality text, translating languages, writing different kinds of creative content, and answering your questions in an informative way. LLMs are pushing the boundaries of natural language processing and are opening up new possibilities for human-computer interaction.
37. **Latent Space:** A low-dimensional representation of data learned by a machine learning model. This compressed representation captures the essential features and relationships within the data, allowing the model to efficiently perform tasks such as image generation, translation, and anomaly detection.
38. **Latent diffusion**: Latent diffusion is a generative modeling technique that uses a diffusion process to gradually add noise to a latent representation of the data. By reversing this process, the model can learn to denoise the data and generate new samples that are similar to the training data. Latent diffusion models have been shown to be effective for generating high-quality images, text, and other types of data.
39. **LLamaIndex:** An indexing method specifically designed for large language models (LLMs). It improves the retrieval capabilities of LLMs by efficiently searching through large amounts of text and identifying relevant passages based on the user’s query.

40. **Langchain:** A framework for chaining together different language models with complementary capabilities. This enables the creation of more complex and versatile language processing systems that can handle diverse tasks requiring different skills and knowledge.

41. **LLMOps:** LLMOps (Large Language Model Operations) is a term used to describe the practices and tools involved in developing, deploying, and managing large language models. This includes tasks such as model training, inference, monitoring, and governance. As LLMs become increasingly complex and are used in more mission-critical applications, LLMOps is becoming an increasingly important area of focus.

42. **LoRA:** LoRA (Low-Rank Adapter) is a technique for adapting large language models (LLMs) to specific downstream tasks with minimal fine-tuning. It works by training a small adapter module on the specific task data, which is then plugged into the pre-trained LLM. This approach can significantly reduce the amount of training data and computational resources required for fine-tuning, while still achieving good performance.

M

43. **Mixture of Experts:** A machine learning ensemble method that combines the predictions of multiple, specialized submodels to improve overall performance. Each submodel is trained on a specific aspect of the problem, and their predictions are aggregated to produce a final prediction.

44. **Multimodal AI:** Machine learning models that can process and generate data from different modalities, such as text, images, audio, and sensor data. This allows them to understand and respond to the world in a more comprehensive way, with applications in areas such as robotics, healthcare, and the internet of things.

N

45. **NeRF (Neural Radiance Fields):** A novel method for creating 3D scenes from 2D images. NeRF models represent a 3D scene as a continuous function that predicts the color and density of light passing through each point in space. This allows them to generate photorealistic images from any viewpoint, even if the viewpoint was not included in the original training data.

O

46. **Objective Function:** A function that is maximized or minimized during the training of a machine learning model. The choice of the objective function determines what the model is trying to learn and how it measures its success. Different tasks may have different objective functions, such as minimizing the error in predictions for regression tasks or maximizing the likelihood of the data for classification tasks.
47. **One-Shot Learning**: This approach enables a machine learning model to learn from only one example per class. It's crucial for applications where data is scarce, allowing the model to make predictions or recognize patterns based on a single instance.

P

48. **PEFT (Prompt Engineering Fine-Tuning):** A method for enhancing the performance of large language models through tailored prompt engineering and fine-tuning. Prompt engineering involves carefully crafting the instructions and examples given to the model, while fine-tuning involves retraining the model on a small dataset of data tailored to the specific task

49. **Pre-training:** The kindergarten of AI models, where they learn fundamental skills like recognizing patterns and extracting features from data. Think of it as building a vocabulary before writing a story.

50. **Prompt:** The question, riddle, or instruction that kicks off an AI model's task. It sets the direction and context for what the model should generate or predict, like prompting a writer with a theme and genre.

51. **ProGAN (Progressive Growing of GANs):** The artistic AI chef, starting with rough sketches and gradually adding details until stunningly realistic images emerge. It uses Generative Adversarial Networks (GANs) in a step-by-step process to refine output.

Q

52. **QLoRA**: QLoRA (Quantized LoRA) is a further refinement of LoRA that uses quantization to reduce the size and memory footprint of the adapter module. This makes it possible to deploy LLMs on devices with limited resources, such as mobile phones or edge devices.

R

53. **Regularization:** The AI gym coach, preventing models from overfitting or memorizing specific training data. It uses techniques like adding constraints or noise to encourage flexibility and generalizability, allowing the model to perform well on unseen examples.

54. **Reinforcement Learning (RL):** An iterative learning paradigm where an agent interacts with an environment to maximize a reward signal. Through trial and error, the agent learns to map actions to optimal outcomes, excelling in complex tasks like robotics and game playing.

55. **RLHF (RL from Human Feedback):** Accelerates RL by directly incorporating human expertise in the form of rewards, penalties, or demonstrations. This allows for faster learning and refinement of the agent's policy, particularly in situations with challenging reward functions.

S

56. **Self-Supervised Learning (SSL):** Exploits unlabeled data by generating its own labels from inherent patterns and structures. SSL leverages diverse techniques like contrastive learning or inpainting to achieve remarkable results in image recognition, NLP, and speech recognition.
57. **Sequence-to-Sequence Models (Seq2Seq)**: Models that transform a sequence of elements (like words in a sentence) into another sequence. They are vital in applications like machine translation and speech recognition.
58. **StyleGAN:** A family of GANs specialized in generating highly realistic and customizable human faces. StyleGAN utilizes a two-stage approach: capturing facial features in a latent space and progressively refining the image, allowing for style manipulation through additional control parameters.
59. **Singularity:** A hypothetical point in time where technological advancement, particularly in AI, surpasses human control and understanding. The Singularity remains a speculative and debated concept, but it raises crucial questions about the future of technology and its potential impact on humanity.

T

60. **Text-to-Speech (TTS):** A subfield of NLP focused on converting written text into spoken voice output. TTS leverages statistical parametric synthesis and deep learning techniques to generate realistic and expressive speech, impacting sectors like screen readers, voice assistants, and text-to-audio applications.

61. **TPU (Tensor Processing Unit):** A specialized microprocessor designed by Google for AI workloads. TPUs are optimized for the highly parallel computations required for training and running large neural networks, offering significant performance and efficiency gains compared to traditional CPUs.

62. **Transfer Learning:** A technique for leveraging knowledge acquired from a pre-trained model to solve new problems. This reduces the amount of training data and time required compared to training a model from scratch, making it widely used in image recognition, NLP, and robotics.

63. **Transformer:** A type of neural network architecture specifically designed for processing sequential data like text and code. Transformers utilize self-attention mechanisms to capture long-range dependencies within the data, leading to state-of-the-art performance in various NLP tasks.

V

64. **Variational Autoencoders (VAEs):** A type of generative model that uses neural networks to encode data into a lower-dimensional space and then reconstruct it. VAEs are crucial for tasks like image generation and anomaly detection.

65. **Vector Databases:** Vector databases are specialized databases designed to store and efficiently query high-dimensional vectors. These vectors represent data points in a multidimensional space, such as

text documents, images, or time series data. Vector databases use indexing and search techniques optimized for vector data to enable fast retrieval of similar vectors or vectors that satisfy specific criteria.

X

66. **XAI (Explainable AI):** A research field aimed at making AI models more interpretable and understandable. XAI techniques help explain how models make decisions, identify potential biases, and build trust between humans and AI systems.

Z

67. **Zero-shot Learning:** In zero-shot learning, a model is capable of handling tasks it has never explicitly been trained on. It leverages understanding from different but related data or tasks to infer completely new categories, making it useful when training data for specific tasks is unavailable.