COURANT AI™
Inventive Intelligence

AI-100: Terminology & Concepts

An introduction to artificial intelligence basic terms and topics.

1-1-25 Version 1.2.0

Brief History of AI

Artificial Intelligence (AI) as a field has its roots in the mid-20th century, emerging from a blend of computer science, mathematics, psychology, linguistics, and philosophy. Its development has come in waves, influenced by technological capabilities, funding, and changing societal expectations. Understanding the history of AI helps contextualize the advancements leading to models like ChatGPT that we commonly use today.

Early Foundations (1940s–1950s)

Conceptual Origins: Even before electronic computers, visionaries like Alan Turing pondered whether machines could think. Turing's 1950 paper, "Computing Machinery and Intelligence," introduced the famous Turing Test as a criterion for machine intelligence.

Logic and Symbolic Reasoning: Researchers explored how to represent human thought processes as logical rules that computers could follow. This marked a shift from purely numerical computation to symbolic manipulation.

The Dartmouth Conference and the Birth of AI (1956–1960s)

Dartmouth Workshop (1956): Often cited as AI's founding event, a small group of scholars—including John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon—met at Dartmouth College to discuss the potential of machines that "use language, form abstractions and concepts, solve kinds of problems now reserved for humans."[1]

Early Optimism: This period saw a wave of optimism. Programs like the Logic Theorist (1956) and later the General Problem Solver (GPS) demonstrated that computers could solve puzzles and prove mathematical theorems. AI researchers believed human-level intelligence in machines might be achieved in a matter of decades.

Expert Systems and Symbolic AI (1960s–1970s)

Knowledge Representation: AI research in the 1960s focused on how to represent knowledge in symbolic structures (e.g., semantic networks, frames). Computers attempted to reason by manipulating these representations.

Expert Systems: In the 1970s and early 1980s, expert systems like MYCIN (a medical diagnosis system) encoded specialized human knowledge into rules. They achieved success in narrow domains, prompting industry interest and fueling optimism that AI could revolutionize business and science as well as the daily lives of thinking humans.

First AI Winter (1970s–1980s)

Unmet Expectations: The grand claims of early AI research proved difficult to realize, especially in natural language understanding and general reasoning. Computing power was limited, and the complexity of real-world problems overwhelmed symbolic approaches.

Funding Cuts and Disillusionment: Governments and industry sponsors, disappointed with slow progress, reduced funding. The field entered its first "AI Winter," a period of decreased momentum and low public enthusiasm.

The Rise of Machine Learning (1980s–1990s)

Shift from Rules to Data: AI research began moving away from handcrafted rules toward machine learning—techniques that allowed computers to learn patterns from data.

Neural Networks Revival: Although neural networks had been theorized as early as the 1940s and 1950s (e.g., McCulloch-Pitts neuron model, Rosenblatt's Perceptron), they fell out of favor due to theoretical limitations. In the 1980s, the backpropagation algorithm revitalized interest, enabling multi-layer networks to learn more complex tasks.

Probabilistic Reasoning and Statistics: Researchers incorporated statistical methods for dealing with uncertainty, giving rise to Bayesian networks and Markov models. This era positioned AI closer to what we now call machine learning and statistical inference.

Second AI Winter and Modest Resurgence (Late 1980s–1990s)

Overhyped Expectations Again: AI went through another slowdown as some applications failed to meet commercial expectations.

Steady Progress in Research: Despite funding challenges, researchers continued refining ML algorithms, exploring decision trees, support vector machines, and ensemble methods.

The Age of Big Data and Deep Learning (2000s–2010s)

Computational Power and Data Growth: By the mid-2000s, faster CPUs, GPUs, and the internet's explosion of digital content provided a foundation for training more complex models.

Deep Learning Breakthroughs: In 2012, a deep convolutional neural network (AlexNet) dramatically improved image recognition performance at the ImageNet competition. This breakthrough ignited a deep learning revolution, with neural networks quickly surpassing older machine learning techniques in computer vision, speech recognition, and natural language processing.

Infrastructure and Tools: Open-source libraries like TensorFlow and PyTorch simplified building and training neural networks. Cloud computing platforms made it easier to access massive computational resources.

Natural Language Processing (NLP) and Transformers (2017–2020)

The emergence of Word Embeddings and Recurrent Networks: Before deep learning fully took hold, techniques like word embeddings (e.g., Word2Vec) and recurrent neural networks (RNNs) significantly improved the quality of machine translation and text understanding.

Transformers and Self-Attention: The 2017 paper "Attention Is All You Need" introduced the Transformer architecture, revolutionizing NLP. Transformers rely on attention mechanisms rather than recurrence, making it easier to process long-range dependencies in text.

Pretrained Language Models: With the Transformer architecture, researchers developed large-scale language models trained on vast amounts of text. Models like BERT (2018) excelled in various NLP tasks by providing pre-trained "understandings" of language that could be fine-tuned for specific tasks.

Scaling Up and the Advent of Generative AI (2020–2022)

Bigger Models, Better Results: As researchers scaled up the size of neural networks and training datasets, models like GPT-2 and GPT-3 (Generative Pretrained Transformers) demonstrated remarkable abilities in generating coherent, contextually relevant text. GPT-3, released in 2020, highlighted that general-purpose language models could perform tasks they weren't explicitly trained for, from writing essays to answering coding questions.

Multimodal AI: Concurrently, developments in combining text with images, audio, or other modalities expanded AI's capabilities, enabling systems like CLIP and DALL·E to generate images from text prompts or understand complex visual scenes.

Toward ChatGPT (2022)

Conversations with AI: Although large language models like GPT-3 could generate impressive text, they lacked the refined ability to hold a coherent, helpful conversation reliably and safely. Researchers began focusing on instruction following and dialogue fine-tuning.

Instruction-Tuning and Reinforcement Learning from Human Feedback (RLHF): By training large language models with instructions and human feedback, developers aligned the models' outputs more closely with human expectations, correctness, and helpfulness. This process improved the model's ability to interact conversationally and avoid problematic outputs.

ChatGPT Debuts (Late 2022): ChatGPT emerged as a user-friendly interface on top of large language models fine-tuned for dialogue. It demonstrated an unprecedented ability to understand complex questions, provide detailed explanations, and even engage in creative content generation. With its release, AI became *more accessible to the general public*, sparking widespread discussions about the future of work, education, ethics, and trust in AI.

In Summary

Artificial Intelligence began as an ambitious idea that machines could mimic human reasoning. Early systems relied on handcrafted rules and struggled under the weight of real-world complexity. The field evolved with statistical methods, machine learning, and then deep learning, propelled by increasing computational power and data availability.

The development of neural networks, especially with the Transformer architecture, led to the creation of powerful large language models. By the time ChatGPT emerged, AI had transformed from an academic pursuit into a practical tool impacting numerous fields—demonstrating that, while the path was neither direct nor without setbacks, AI's trajectory consistently moved toward more capable and human-like systems.

Types of AI

Based on *Methodologies* (there are several ways you can organize types of AI). This classification refers to the techniques and approaches used to implement AI.

Rule-Based Systems

Definition: AI that relies on predefined rules.
Examples: Expert systems, decision trees.
Characteristics: Deterministic and explicit logic.

Machine Learning (ML)

Definition: AI that learns patterns and makes decisions based on data.
Subcategories:

- Supervised Learning: Labeled data for training.
- Unsupervised Learning: Patterns from unlabeled data.
- Reinforcement Learning: Learning through rewards and penalties.

Examples: Image recognition, fraud detection.
Generative AI is a subset of Machine Learning that focuses on creating new data based on learned patterns from existing datasets.

Deep Learning

Definition: A subset of ML using neural networks with many layers.
Examples: Speech recognition, natural language processing.
Characteristics: High computational power, data-intensive.
Generative AI employs advanced Deep Learning techniques, such as:

- *Transformers: Used in text generation models like GPT.*
- *Generative Adversarial Networks (GANs): Used in image generation.*
- *Diffusion Models: Used for high-quality visual content creation.*

Natural Language Processing (NLP)

Definition: AI focused on understanding and generating human language.
Examples: Chatbots, translation systems.
Characteristics: Text and speech-based AI applications.
When applied to tasks like text generation, conversational AI, or translation, Generative AI aligns with Natural Language Processing. Example: Generative AI models like GPT-4 or Claude are designed specifically for understanding and producing human language, making them a central part of NLP.

Robotics

Definition: AI integrated into robots to perform physical tasks.
Examples: Industrial robots, robotic assistants.
Characteristics: Combines AI with physical systems.

Computer Vision

Definition: AI focused on interpreting visual information.
Examples: Facial recognition, object detection.
Characteristics: Vision-based tasks.

Evolutionary Algorithms

Definition: AI inspired by biological evolution, optimizing solutions iteratively.
Examples: Genetic algorithms.
Characteristics: Problem-solving via natural selection principles.

Basic Terms

These initial terms provide a solid foundation for beginners by covering broad concepts and fundamental building blocks of AI. These terms help newcomers understand what AI is, how it works, and key processes involved in training and evaluating models.

Alignment - The idea is that AI systems should align with human values and follow human intent, not do their own thing and go rogue. A big part of this is a process called reinforcement learning with human feedback (RLHF).

Algorithm - A set of step-by-step instructions that a computer follows to perform a specific task or solve a problem.

Artificial Intelligence (AI) - The field of computer science focused on creating machines capable of performing tasks that typically require human intelligence, such as reasoning, learning, or problem-solving.

Bias - Systematic errors or unfair tendencies in machine learning models resulting from biased data, assumptions, or training processes.

Big Data - Extremely large and complex data sets that cannot be easily handled by traditional tools, often used to uncover patterns or insights through advanced analytics and AI techniques.

Computer Vision - An area of AI that allows machines to interpret and understand visual information from images or videos, making it possible to recognize objects, faces, or scenes.

CPU - Central processing unit of a computer or server. The primary processing unit of a computer responsible for executing general-purpose instructions, including some machine learning tasks. Less optimized for large-scale parallel computations compared to GPUs.

Deep Learning - A specialized area of machine learning that uses multi-layered neural networks to automatically learn representations and patterns from large amounts of data.

Ethics - The principles and guidelines that ensure AI systems are developed and used responsibly, fairly, and in a way that respects societal values and human rights.

Feature Extraction - The process of selecting and transforming raw data into a set of relevant characteristics (features) that a model can use to make better predictions or decisions.

Generalization - The ability of a machine learning model to perform well on new, unseen data rather than just memorizing details from the training data.

Generative AI - AI methods, such as generative adversarial networks (GANs), that create new content—such as images, text, or audio—resembling the data on which they were trained.

GPU - Graphical processing unit of a computer or server. A specialized processor designed for high-performance parallel computations, making it highly efficient for training and running machine learning models, particularly those involving large-scale matrix operations.

Inference - The process of applying a trained model to new, unseen data to make predictions or draw conclusions.

Machine Learning (ML) - A subset of AI that enables systems to learn patterns from data and improve their performance over time without being explicitly programmed for each task.

Model - A structured representation (often a set of mathematical relationships) that a machine learning algorithm creates and uses to make predictions or decisions based on input data.

Natural Language Processing (NLP) - A branch of AI focused on enabling computers to understand, interpret, and generate human language, both written and spoken.

Neural Network - A computing structure inspired by the human brain's network of neurons, consisting of layers of interconnected nodes that process information and learn patterns through weights and adjustments.

Overfitting - A situation where a model learns the training data too well, including noise and irrelevant details, causing it to perform poorly on new, unseen data.

Parameters - The number of parameters typically represents the size and complexity of the model. These parameters are responsible for encoding linguistic knowledge and patterns derived from the training data. [See more detail below.]

Reinforcement Learning - A type of machine learning where an agent learns by interacting with an environment, receiving rewards or penalties for its actions, and adjusting its behavior to maximize long-term gains.

Supervised Learning - A type of machine learning where the model is trained on labeled data (with known correct answers) so it can make predictions on new, unseen examples.

Tokens - LLMs like Chat-GPT are trained on all books and the entire internet, but instead of words, AI models break text down into tokens. Many words map to single tokens, though longer or more complex words often break down into multiple tokens. GPT-3 was trained on roughly 500 billion tokens.

Training Data - The examples (inputs and outputs) used to teach a machine learning model how to perform a desired task by recognizing patterns.

Unsupervised Learning - A type of machine learning where the model identifies patterns and structures within unlabeled data, discovering hidden relationships or groupings.

Advanced Terms

The next set of terms introduce more advanced or specialized concepts. Learners will find these terms to be helpful after they grasp the basics because they will augment your understanding and demonstrate the evolving nature of AI technologies.

Backpropagation - The process of calculating and distributing errors back through a neural network, enabling the network to adjust its weights to reduce future errors.

Bayesian Networks - Graphical models that represent probabilistic relationships among variables using nodes for variables and directed edges for dependencies, enabling reasoning under uncertainty.

Convolutional Neural Network (CNN) - A type of neural network primarily used for processing and analyzing visual data, leveraging filters (kernels) to detect features like edges and shapes.

Data Augmentation - Techniques for increasing the variety and quantity of training data (e.g., rotating images, paraphrasing text) to improve model robustness.

Data Labeling - The process of assigning meaningful tags or categories to raw data (images, text) so that a model can learn to recognize these labels.

Embeddings - Numerical representations of complex data (like words, images) in a continuous vector space, enabling machines to understand semantic relationships.

Ensemble Methods - Approaches that combine multiple models to produce more accurate and stable predictions than any single model alone.

Evaluation Metrics - Measures (such as accuracy, precision, recall) used to assess a model's performance and compare different models or approaches.

Explainable AI (XAI) - Techniques and tools aimed at making AI decisions more transparent and understandable, helping users trust and interpret model outputs.

Feature Engineering - The process of selecting, transforming, and creating relevant input variables (features) that improve a model's predictive power.

Federated Learning - A decentralized approach to training models directly on devices (e.g., smartphones) so that data remains local, enhancing privacy and efficiency.

Gradient Descent - An optimization method used to adjust model parameters step-by-step, moving them closer to the set of values that minimize the model's error.

Hyperparameter - Configuration settings (not learned by the model) that influence how a machine learning algorithm trains and performs, such as learning rate or number of layers.

Instruction-Tuning - The process of fine-tuning a pretrained model using examples of task-specific instructions and responses to improve its ability to follow natural language directives.

Large Language Models (LLMs): AI models trained on vast amounts of text data, capable of generating human-like text and understanding complex language patterns.

Loss Function - A measure of how far a model's predictions are from the target values, guiding the training process to improve accuracy.

Open Source Libraries - Open-source libraries are collections of reusable code made freely available under open-source licenses, allowing developers to use, modify, and distribute them in their own projects.

Markov Models - Mathematical frameworks that represent systems where future states depend only on the current state and not on the sequence of past states, enabling probabilistic modeling of sequential processes.

Pretrained Language Models - Neural networks trained on large text corpora to understand and generate human language, serving as a foundation for fine-tuning on specific tasks.

Probabilistic Reasoning and Statistics - Probabilistic reasoning and statistics involve using mathematical frameworks and statistical methods to model uncertainty, infer patterns, and make predictions based on incomplete or noisy data.

Prompt Engineering - The practice of designing and refining the instructions (prompts) given to language models to produce more accurate, relevant, or creative outputs.

Recurrent Neural Network (RNN) - A neural network designed to handle sequential data (e.g., time series, text) by maintaining an internal state that captures information from previous inputs.

Regularization - Techniques designed to prevent overfitting by penalizing overly complex models, helping them generalize better to unseen data.

Reinforcement Learning from Human Feedback - A training approach where a model learns to align its behavior with human preferences by optimizing rewards based on evaluations provided by human feedback.

Robotics - The integration of AI into machines that can sense, plan, and execute tasks in the physical world, enabling automated manufacturing, delivery, or exploration.

Self-Attention - A mechanism that allows a model to dynamically focus on different parts of an input sequence to understand contextual relationships and dependencies between elements, regardless of their distance in the sequence.

Transfer Learning - A technique where a model trained on one task is repurposed and refined for another related task, reducing training time and improving performance with limited data.

Transformer: A neural network architecture that processes input data in parallel rather than sequentially, allowing for more efficient handling of long-range dependencies in language tasks.

Leaders in AI

Below are ten prominent AI companies recognized for their significant contributions, research, products, and services. The landscape is rapidly evolving, but these organizations are generally regarded as leaders in AI as of today.

OpenAI - Main Products/Services: GPT-series language models (e.g., GPT-4), ChatGPT conversational AI system, DALL·E for image generation, and APIs allowing developers to integrate advanced AI into their applications.

Google (Alphabet) - Main Products/Services: DeepMind research division, the Bard conversational AI, Google Cloud's Vertex AI platform, TensorFlow (an open-source AI library), and AI-driven features integrated into Search, Gmail, and Google Workspace.

Microsoft - Main Products/Services: Azure Cognitive Services and Azure Machine Learning, GitHub Copilot (AI code assistant), and integrated AI capabilities across the Microsoft 365 suite. Microsoft also has a major investment and partnership with OpenAI.

Amazon (AWS) - Main Products/Services: Amazon SageMaker for building and training machine learning models, AWS AI services (such as Amazon Rekognition for image analysis and Amazon Comprehend for NLP), and customized AI solutions for businesses at scale. Amazon Q for generative AI.

NVIDIA - Main Products/Services: High-performance GPUs optimized for machine learning and deep learning workloads, CUDA platform, AI frameworks like NVIDIA Triton Inference Server, and NVIDIA DGX systems for enterprise AI infrastructure.

Meta (Facebook) - Main Products/Services: Development and open-sourcing of PyTorch (a leading deep learning framework), advanced AI research in language models (e.g., LLaMA), computer vision, and content moderation. AI also powers feeds, advertising, and VR/AR initiatives across Meta's platforms.

IBM - Main Products/Services: Watson AI solutions for enterprise, Watsonx AI and data platform, and a suite of AI-driven products focused on business optimization, natural language understanding, and predictive analytics in fields like healthcare, finance, and customer service.

Salesforce - Main Products/Services: Einstein AI (integrated AI layer within Salesforce CRM platform), generative AI for sales and marketing insights, and predictive analytics tools that help businesses improve customer engagement and decision-making.

Baidu - Main Products/Services: ERNIE language models, Baidu Brain (an AI platform offering NLP, vision, and speech services), autonomous driving platform Apollo, and AI tools integrated into search, advertising, and cloud services primarily serving the Chinese market.

Anthropic - Main Products/Services: Claude (a large language model focused on safety, interpretability, and controllability), enterprise partnerships for AI integration, and research aimed at building safer, more responsible large-scale AI systems.

Apple - Key Contributions: On-device machine learning technologies integrated into products like the iPhone and Apple Watch (e.g., Face ID, Siri's voice recognition), Core ML framework for app developers, and privacy-focused AI research.

Hugging Face - Key Contributions: A leading open-source repository and platform for sharing and collaborating on AI models, including the popular Transformers library. They foster a community-driven approach, hosting a wide array of ready-to-use models and datasets for NLP, vision, and more.

Intel - Key Contributions: AI-optimized hardware (CPUs, VPUs, and Nervana Neural Network Processors), software toolkits for machine learning (OpenVINO), and investments in edge computing solutions to enable efficient, low-latency inference on a wide range of devices.

Tesla - Key Contributions: Pioneering autonomous driving technology powered by computer vision and deep learning, developing custom AI chips for in-vehicle neural networks, and training large-scale models to improve driver assistance and full self-driving capabilities.

Adobe - Key Contributions: Adobe Sensei, an AI and machine learning framework integrated into Adobe's creative tools and marketing platforms. It enhances image editing, automates design suggestions, personalizes marketing campaigns, and accelerates workflow productivity for creators and businesses.

C3 AI - Key Contributions: Enterprise AI software for predictive analytics, supply chain optimization, fraud detection, and customer relationship management. C3 AI's platform helps businesses integrate, manage, and analyze large datasets to drive decision-making.

Palantir - Key Contributions: AI-powered data integration and analytics platforms like Palantir Foundry and Gotham. They help organizations (particularly large enterprises and governments) uncover insights from complex, disparate data sources, aiding strategic and operational decision-making.

Graphcore - Key Contributions: Design and manufacture of Intelligence Processing Units (IPUs), specialized chips that accelerate machine learning workloads. Graphcore's hardware innovations enable more efficient training and inference for large, complex AI models.

Boston Dynamics - Key Contributions: Advanced robotics integrated with AI-driven perception and control systems. Their robots (e.g., Spot and Atlas) showcase sophisticated AI for navigation, object recognition, and agility in dynamic, real-world environments.

Preferred Networks - Key Contributions: A Japanese AI startup focused on applying deep learning to areas like manufacturing, robotics, healthcare, and IoT. They've contributed to advancements in reinforcement learning, autonomous vehicles, and industrial process optimization, often partnering with established corporations.

These leaders play critical roles across various segments of the AI landscape, from hardware innovation and open-source platforms to specialized enterprise solutions and cutting-edge robotics. They drive innovation through foundational research, cloud platforms, applied machine learning models, and integrated AI solutions affecting many industries worldwide.

Leading LLMs

As of early 2025, the landscape of large language models (LLMs) has expanded significantly, with several models gaining prominence due to their capabilities and widespread adoption. Here are ten of the most popular LLMs.

GPT-4o: Developed by OpenAI, GPT-4o is the latest iteration in the GPT series, offering enhanced performance and multimodal capabilities, including text, image, video, and voice processing.

Claude 3.5: Anthropic's Claude 3.5 Sonnet is a significant upgrade from its predecessors, featuring a 200,000-token context window and improved performance in coding and text reasoning tasks.

Grok-1: Released by xAI, Grok-1 is notable for its 314 billion parameters and integration with the X (formerly Twitter) platform, offering users advanced conversational AI capabilities.

Mistral 7B: An open-source model from Mistral AI, Mistral 7B has 7.3 billion parameters and outperforms many larger models in tasks like reading comprehension and coding.

PaLM 2: Google's PaLM 2 is an advanced LLM trained on 3.6 trillion tokens, excelling in reasoning, question answering, and coding tasks.

Falcon 180B: Developed by the Technology Innovation Institute, Falcon 180B is an open-source model with 180 billion parameters, outperforming models like GPT-3.5 and LLaMA 2 in various tasks.

Stable LM 2: From Stability AI, Stable LM 2 includes models with 1.6 billion and 12 billion parameters, offering strong performance in language understanding and generation.

Gemini 1.5: Google DeepMind's Gemini 1.5 provides a significant upgrade over its predecessor, featuring a one million-token context window, the largest to date among LLMs.

Llama 3.1: Meta AI's Llama 3.1 boasts 405 billion parameters and a 128,000-token context length, making it one of the largest open-source models available.

Mixtral 8x22B: Mistral AI's Mixtral 8x22B is a sparse Mixture-of-Experts model with 141 billion parameters, designed to optimize performance and cost efficiency.

These models represent the forefront of LLM development, each offering unique features and capabilities that cater to a wide range of applications in natural language processing and understanding.

Ollama for Local LLMs

Ollama is a robust platform designed to facilitate the local deployment of large language models (LLMs). It simplifies the process of running open-source LLMs on personal computers, offering several advantages.

Key Features of Ollama

- Local Model Execution: Ollama enables users to run LLMs directly on their local machines, enhancing data privacy and allowing for offline usage.
- Open-Source Model Support: The platform is compatible with various open-source models, providing transparency and flexibility for users to inspect and modify models as needed.
- User-Friendly Setup: Ollama offers an easy installation and configuration process, making it accessible to users with varying levels of technical expertise.
- Customization Options: Users can fine-tune model performance through Modelfiles, adjusting parameters and integrating additional data to optimize models for specific use cases.
- Developer API: Ollama provides a robust API that developers can leverage to integrate AI functionalities into their software, supporting various programming languages and frameworks.

Benefits of Using Ollama for Local LLM Deployment

There are several benefits to leveraging Ollama for local LLM deployment and development.

- Data Privacy: Running models locally ensures that sensitive data remains on your machine, reducing exposure risks.
- Performance Efficiency: Local processing can offer faster response times, especially for frequent queries, as it eliminates reliance on external servers.
- Cost Savings: By avoiding cloud service fees, users can reduce costs associated with deploying and running LLMs.
- Offline Accessibility: Once models are installed, they can operate without an internet connection, ensuring continuous availability.

Ollama Considerations

While Ollama offers numerous advantages, it's important to consider the hardware requirements for running LLMs locally. Larger models may necessitate significant computational resources, including substantial RAM and, in some cases, high-performance GPUs. Additionally, users are responsible for managing updates and ensuring that models and software remain current.

In summary, Ollama is a valuable tool for those looking to deploy LLMs locally, providing a balance between ease of use, customization, and control over data privacy.

Meta LLaMa

The size of Meta's LLaMA (Large Language Model Meta AI) models varies depending on the specific version and number of parameters. Here's a general breakdown for installing them locally.

LLaMA Model Sizes

Including approximate disk space requirements.

1. LLaMA 7B: Around 13-15 GB (depending on the quantization and model type).
2. LLaMA 13B: Around 26-30 GB.
3. LLaMA 30B: Around 60-70 GB.
4. LLaMA 65B: Around 130-150 GB.

Installation Considerations

- Quantization: You can significantly reduce the model size by using quantized versions (e.g., 4-bit or 8-bit). Quantized models are smaller and often faster but may lose some precision.
    - Example: A quantized 65B model might only require ~50 GB.
- Hardware Requirements: The larger models (30B and 65B) may require high-performance GPUs with sufficient VRAM (e.g., 24GB+ for 30B, and 48GB+ for 65B).
- RAM and Storage: Ensure you have enough system RAM and SSD storage for smooth operation. Generally, you need at least twice the model size in RAM.

If you're aiming for local deployment, ensure your hardware aligns with the model's size and requirements.

Parameters

In the context of machine learning and large language models (LLMs), a parameter refers to a numeric value within the model that is learned during the training process. Parameters are the building blocks of a neural network and determine how the model processes input data to generate output. Here's a detailed explanation:

What Are Parameters?

Parameters are the weights and biases within a model's architecture. These values are adjusted during training to minimize the error between the model's predictions and the actual target values.

- **Weights**: Represent the strength of the connection between neurons in adjacent layers of the model.
- **Biases**: Allow the model to shift the activation function and make predictions more flexible.

Parameters in LLMs

In LLMs like GPT, LLaMA, or Mistral:

- The number of parameters typically represents the size and complexity of the model.

- These parameters are responsible for encoding linguistic knowledge and patterns derived from the training data.

For example:
- GPT-3 has 175 billion parameters.
- LLaMA 2 13B has 13 billion parameters.

More parameters usually mean a more powerful model but require:
- More computational resources to train and run.
- Larger datasets for training.

What Parameters Do

Parameters determine several factors.
1. Determine Model Behavior: Parameters define how the model generates responses, translates languages, or solves tasks.
2. Store Learned Knowledge: During training, the model adjusts parameters to "learn" from data.
3. Impact Model Size: The total number of parameters is proportional to the storage and memory requirements of the model.

Trade-offs of Parameter Count

More Parameters
- **Pros**: Higher capacity to learn complex patterns; but also an incrementally better performance on diverse tasks.
- **Cons**: Increased computational cost, latency, and memory requirements.

Fewer Parameters
- **Pros**: Faster inference, reduced hardware requirements.
- **Cons**: Limited ability to handle complex tasks or generalize.

Simplified Analogy

Think of parameters as the dials and knobs on a machine. Training a model is like fine-tuning these knobs to make the machine perform optimally for specific tasks. Each parameter influences how the model processes input data and generates output.

Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and humans through natural language. The goal of NLP is to

enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP involves several key tasks.

Text Analysis: Breaking down and understanding text data, including tasks like tokenization (splitting text into words or phrases), part-of-speech tagging, and syntactic parsing.

Sentiment Analysis: Determining the emotional tone behind a piece of text, such as identifying whether a movie review is positive or negative.

Machine Translation: Automatically translating text from one language to another, as seen in services like Google Translate.

Speech Recognition: Converting spoken language into text, used in applications like voice-activated assistants (e.g., Siri or Alexa).

Text Generation: Creating new text based on a given input, which is a core function of generative AI models like GPT.

NLP is foundational for many applications, from chatbots and virtual assistants to automated summarization and language translation services. It bridges the gap between human communication and computer understanding.

Machine Learning

Machine Learning (ML) is a broader field within artificial intelligence focused on developing algorithms and models that allow computers to learn from and make predictions or decisions based on data. Unlike traditional programming, where explicit instructions are given to perform tasks, machine learning systems improve their performance on a task by identifying patterns in data without being explicitly programmed for each specific task.

Machine Learning (ML) vs. Natural Language Processing (NLP)

Here are some points of comparison on how Machine Learning differs from Natural Language Processing (NLP).

Scope

- **Machine Learning**: Encompasses a wide range of tasks and applications beyond language, including image recognition, fraud detection, recommendation systems, and more. It's a general-purpose approach to making predictions or decisions based on data.
- **NLP**: A specific subfield of AI and machine learning focused on enabling computers to understand, interpret, and generate human language.

Applications

- **Machine Learning**: Used in various domains like finance (for credit scoring), healthcare (for disease prediction), and marketing (for customer segmentation).
- **NLP**: Primarily used for tasks involving text or speech, such as sentiment analysis, machine translation, speech recognition, and text generation.

Techniques

- **Machine Learning**: Involves a variety of techniques including supervised learning, unsupervised learning, reinforcement learning, and deep learning. These techniques are applied to structured data, images, audio, and more.
- **NLP**: Uses machine learning techniques, particularly deep learning, to analyze and process language data. It often involves specialized models like transformers, which are particularly suited to handling sequences of words or characters.

In essence, while machine learning is a broad field concerned with creating systems that can learn from data, NLP is a specific application of machine learning focused on enabling machines to process and understand human language.

Deep Learning

Deep Learning is a subset of machine learning that involves Neural Networks with many layers (hence “deep”) to model and understand complex patterns in large amounts of data. It’s particularly effective in handling unstructured data like images, audio, and text. There are several key characteristics of deep learning.

Neural Networks: Deep learning models are based on artificial neural networks, which are inspired by the structure and function of the human brain. These networks consist of layers of interconnected nodes (neurons) that process input data to make predictions or decisions.

Multiple Layers: Unlike traditional machine learning models, which might have one or two layers, deep learning models have multiple layers of neurons. These layers allow the model to learn hierarchical features, with each successive layer capturing increasingly abstract information.

For example, in image recognition, early layers might detect edges and corners, while deeper layers might recognize more complex structures like shapes or objects.

Large Datasets and Computational Power: Deep learning models require vast amounts of data and significant computational resources to train effectively. The availability of big data and advancements in hardware, particularly GPUs, have been critical to the success of deep learning.

Applications: Deep learning has been revolutionary in fields like computer vision (e.g., facial recognition, object detection), natural language processing (e.g., language translation, sentiment analysis), and speech recognition. It's also the technology behind generative models like GPT, which can generate human-like text.

In summary, deep learning is a powerful approach within machine learning that uses complex neural networks to automatically discover patterns and representations in data, enabling it to tackle challenging tasks that were previously difficult or impossible for computers.

Fun AI Apps

Below is an introductory collection of links to various AI applications for common tasks.

- chat.openai.com ChatGPT based on OpenAI
- kickresume.com/en Resume creator
- durable.co Website builder
- rose.ai Easy research data
- beautiful.ai Presentations and data visualization
- decktopus.com Presentations
- looka.com Logo and branding
- godmode.space Decision making/problem solving
- gamma.app Beautiful presentation decks
- Auto-GPT is a free, open-source Python application which uses OpenAI's GPT-4 technology.
- unboundcontent.ai Content creation
- orai.com Practice presentations get feedback
- poised.com Practice presentations and get feedback
- monica.im Personal assistant
- agentgpt.reworkd.ai AgentGPT to set up AutoGPTs in your browser

References

1.] The quote comes from the original proposal for the 1956 Dartmouth Summer Research Project on Artificial Intelligence. The citation is typically given as: McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College. Although it was drafted in 1955, the workshop took place in 1956. This proposal is widely recognized as the founding document that set the stage for the field of AI.