Glossary of AI terms

Prompting techniques

- **Prompt Engineering**: The practice of designing effective prompts to guide AI models in generating useful, accurate, or creative responses. Prompt engineering often involves testing different wording, structure, or instructions to optimize outcomes.
  Example: Telling an AI “Explain this like I’m a CFO” vs. “Summarize this in simple terms” leads to different results.

- **Zero shot prompting:** Giving a model a task without providing examples or prior training specific to the task. It relies entirely on the model’s pre-existing knowledge
  Example: Write a catchy headline for a news article about renewable energy.

- **One shot prompting**: Providing the model with one example to demonstrate the desired output, helping it understand the task or style more clearly
  Example: Write a catchy headline for a news article about renewable energy, similar to 'Revolutionary Solar Farm Powers Entire City with Clean Energy.’

- **Few-shot prompting:** Providing the model with multiple examples to demonstrate the pattern, format, or style you want it to follow
  Example: “Write a headline for a news article about renewable energy. Here are two examples:
  - Wind Power Triumphs: Coastal Turbines Meet National Energy Needs.
  - Hydrogen on the Rise: New Fuel Technology Powers Tomorrow’s Transport."

- **Chain of Thought:** A prompting technique that guides the model to reason step-by-step by breaking down complex tasks into smaller, logical steps. This often improves accuracy on reasoning tasks.
  Example: “Write an op-ed about the importance of renewable energy. Think step-by-step:
  1. Generate potential headlines
  2. Outline the article
  3. Write a first draft
  4. Critique the first draft
  5. Write a revised version"

- **Role prompting:** Instructing the AI to adopt a specific role, persona, or perspective when generating its response. This can influence tone, expertise, and style.
  Example: “Take on the role of a scientist who is politically neutral and focuses only on insights supported by data".

- **Jailbreak:** Techniques or prompts designed to bypass an AI system's safety filters or restrictions, causing it to produce outputs it normally wouldn't (such as prohibited content).
  Example: Asking an LLM to roleplay or use coded language to bypass content moderation.

Model types

- **Foundation model:** A large, general-purpose AI model pretrained on vast, diverse datasets across multiple domains. Foundation models can be adapted (fine-tuned) for specific tasks or applications.
  Examples: GPT-4 (text), Llama 2 (text), DALL-E (images).

- **LLM:** A type of foundation model designed to understand and generate human language. LLMs are trained on large-scale text data and can perform tasks such as answering questions, summarizing, and generating content.

- **SLM**: Small Language Models (SLMs) are compact versions of large models, typically optimized for efficiency, lower computational costs, and deployment in constrained environments (like mobile devices). They can be fine-tuned for specific domains but generally trade off some performance for speed and accessibility.

- **Multimodal Model**: An AI model that can process and generate multiple types of data — such as text, images, audio, or video — in a single system. These models enable richer interactions and more complex tasks.

- **Diffusion Models**: A type of generative AI model that creates high-quality images by starting with random noise and gradually refining it into a coherent image. Diffusion models are used in image generation tools like Stable Diffusion and DALL-E 3.

- **Reasoning Model:** AI models designed to solve complex problems by thinking through multiple steps or evaluating different possibilities. Unlike models that rely mainly on pattern recognition, reasoning models apply structured thinking to arrive at better answers.
  Example: A reasoning model asked, "What's the best way to reduce shipping delays?" might consider several options — adding more drivers, changing routes, or adjusting delivery times — before choosing the most effective solution.

Optimizing accuracy

- **Best of N:** A technique where an AI model generates N outputs for the same prompt, and the best one is selected based on a scoring function or criteria. This increases the likelihood of a higher-quality or more accurate response. Best of N enhances reasoning models by giving them more chances to get it right — helpful when reasoning involves uncertainty or complex decisions.

- **Fine-tuning:** Adapting a pretrained model by training it further on a smaller, task-specific dataset. Fine-tuning helps the model specialize in a particular domain or task, such as legal analysis or medical diagnostics.
  Example: Fine-tuning GPT on a set of legal contracts to improve contract analysis.

- **RAG:** A technique where an LLM retrieves relevant external data (from a database or documents) in real time to augment its responses, improving accuracy and reducing hallucinations.

- **Vector Database**: A specialized type of database designed to store and search embeddings (numerical vectors). Vector databases enable fast retrieval of similar content, which is critical for AI tasks like semantic search, recommendation engines, and Retrieval-Augmented Generation (RAG).

- **Embedding**: A numerical representation of data (such as text or images) that captures its meaning and context in a format the AI model can understand. Embeddings allow similar items (words, documents, images) to be compared and grouped based on meaning.
  Example: A customer asking about “invoice processing” might be linked to related topics like “billing” or “accounts payable” using embeddings.

- **Consistency Checking:** A method used to verify whether an AI model’s outputs are reliable and consistent across different prompts, rephrasings, or variations. It helps identify contradictions or hallucinations by comparing answers to the same or similar questions.
  Example: Asking an AI, “What’s the capital of France?” and later, “Which city is the seat of government in France?” — consistency checking ensures the answers match.

- **Hallucination:** When an AI generates false or misleading information, presenting it as if it were true. Hallucinations can include made-up facts, data, or sources.

Model development

- **Pretraining**: The initial phase of training a foundation model on a broad, diverse dataset before it's fine-tuned for specific tasks. Pretraining teaches the model general knowledge and patterns, making it adaptable for many use cases.
  Example: GPT models are pretrained on massive internet text datasets before fine-tuning for chat-specific tasks.

- **Training vs Inference:**
  Training: The process where an AI model learns patterns from data. This step is resource-intensive and done before deployment.
  Inference: The process where a trained model is used to generate answers, predictions, or outputs in real time. Inference is what happens when you interact with a chatbot..

- **Compute**: The processing power (usually measured in FLOPs — floating point operations) required to train or run AI models. "Compute" typically refers to the servers, GPUs, and other hardware needed for AI workloads.

- **Token:** The smallest unit of text processed by a language model, typically representing a word, sub-word, or character. Tokens are also used to measure processing costs and model input/output limits.
  Example: "fantastic" might be split into two tokens: "fan" and "tastic."

- **Reinforcement Learning (RL)**: A machine learning technique where an AI agent learns to make decisions by receiving feedback in the form of rewards or penalties. RL is often used for tasks requiring sequential decision-making.
  Example: Training an AI to play chess by rewarding it for wins and penalizing it for losses.

- **Reinforcement Learning from Human Feedback (RLHF)**: A method where human feedback is used to fine-tune an AI model's behavior by guiding it toward preferred outputs. RLHF is often used to make models safer and more aligned with human expectations.
  Example: OpenAI uses RLHF to improve GPT's helpfulness and reduce harmful outputs by incorporating human judgments into its training.

- **Generative Adversarial Network (GAN)**: A machine learning framework consisting of two neural networks: a generator that creates synthetic data and a discriminator that evaluates whether data is real or generated. The adversarial training process improves the generator's ability to produce realistic outputs. Applications include generating photorealistic images, deepfakes, and art.

Agents

- **Agent**: An AI system that can perform tasks autonomously or semi-autonomously, often by taking actions and interacting with other tools or systems. Agents can chain multiple steps together, manage workflows, and sometimes operate with a degree of "memory."
  Example: An AI agent that processes incoming customer emails, drafts responses, and files support tickets in your CRM.

- **Orchestration**: Coordinating multiple AI models, agents, tools, and data sources to work together in a workflow or application. Orchestration frameworks manage tasks like data retrieval, tool calling, and multi-step reasoning in AI systems.
  Example: LangChain and LlamaIndex are orchestration frameworks that let developers build AI apps combining language models, vector databases, and external tools.

- **Path Planning:** An AI capability typically used in robotics and autonomous systems to determine an optimal route or sequence of actions to reach a specific goal while avoiding obstacles. Increasingly, path planning concepts are applied in AI agents that need to make sequential decisions or navigate complex tasks.
  Example: A warehouse robot uses path planning to move products efficiently, avoiding collisions with shelves and other robots.

- **Memory:** The ability of an AI system — especially an agent — to store and recall information from previous interactions or tasks, enabling it to maintain context over time. Memory allows AI agents to track long-term goals and manage tasks.

Governance & safety

- **Responsible AI (RAI)**: An approach to developing and deploying AI systems that prioritize ethical considerations, fairness, transparency, and accountability. Responsible AI aims to ensure models are safe, respect privacy, and align with human values.

- **AI Alignment**: The process of ensuring an AI system's goals, behaviors, and outputs are aligned with human values, ethical principles, and intended objectives. Alignment reduces the risk of unintended outcomes by keeping the AI system's decisions consistent with human priorities.

- **Red Teaming:** A structured testing process where experts intentionally try to find vulnerabilities, weaknesses, or unsafe behaviors in an AI system. Red teaming helps organizations identify potential risks before deploying AI at scale.

- **Explainability:** The ability to clearly describe how an AI system makes its decisions or predictions. Explainability helps users, regulators, and stakeholders understand why a model produced a specific outcome, increasing trust and making it easier to identify errors or bias.
  Example: An AI loan approval system that explains a rejection by showing the applicant's credit score and income factors.

- **Human-in-the-Loop (HITL)**: A system design where human judgment is integrated into key parts of an AI's decision-making process. HITL ensures oversight, allowing humans to review, approve, or correct AI-generated outputs — especially in high-stakes or complex situations.
  Example: A medical AI suggests diagnoses, but a doctor reviews and confirms the final decision before informing the patient.