Glossary

Fundamental Concepts

Prompt: A prompt is the input, typically in the form of a question, instruction, or statement, that a user provides to an AI model to elicit a response. The quality and structure of the prompt heavily influence the model's output, making prompt engineering a key skill for effectively using AI.

Context Window: The context window is the maximum number of tokens an AI model can process at once, including both the input and its generated output. This fixed size is a critical limitation, as information outside the window is ignored, while larger windows enable more complex conversations and document analysis.

In-Context Learning: In-context learning is an AI's ability to learn a new task from examples provided directly in the prompt, without requiring any retraining. This powerful feature allows a single, general-purpose model to be adapted to countless specific tasks on the fly.

Zero-Shot, One-Shot, & Few-Shot Prompting: These are prompting techniques where a model is given zero, one, or a few examples of a task to guide its response. Providing more examples generally helps the model better understand the user's intent and improves its accuracy for the specific task.

Multimodality: Multimodality is an AI's ability to understand and process information across multiple data types like text, images, and audio. This allows for more versatile and human-like interactions, such as describing an image or answering a spoken question.

Grounding: Grounding is the process of connecting a model's outputs to verifiable, real-world information sources to ensure factual accuracy and reduce hallucinations. This is often achieved with techniques like RAG to make AI systems more trustworthy.

Core AI Model Architectures

Transformers: The Transformer is the foundational neural network architecture for most modern LLMs. Its key innovation is the self-attention mechanism, which efficiently processes long sequences of text and captures complex relationships between words.

Recurrent Neural Network (RNN): The Recurrent Neural Network is a foundational architecture that preceded the Transformer. RNNs process information sequentially, using loops to maintain a "memory" of previous inputs, which made them suitable for tasks like text and speech processing.

Mixture of Experts (MoE): Mixture of Experts is an efficient model architecture where a "router" network dynamically selects a small subset of "expert" networks to handle any given input. This allows models to have a massive number of parameters while keeping computational costs manageable.

Diffusion Models: Diffusion models are generative models that excel at creating high-quality images. They work by adding random noise to data and then training a model to meticulously reverse the process, allowing them to generate novel data from a random starting point.

Mamba: Mamba is a recent AI architecture using a Selective State Space Model (SSM) to process sequences with high efficiency, especially for very long contexts. Its selective mechanism allows it to focus on relevant information while filtering out noise, making it a potential alternative to the Transformer.

The LLM Development Lifecycle

The development of a powerful language model follows a distinct sequence. It begins with Pre-training, where a massive base model is built by training it on a vast dataset of general internet text to learn language, reasoning, and world knowledge. Next is Fine-tuning, a specialization phase where the general model is further trained on smaller, task-specific datasets to adapt its capabilities for a particular purpose. The final stage is Alignment, where the specialized model's behavior is adjusted to ensure its outputs are helpful, harmless, and aligned with human values.

Pre-training Techniques: Pre-training is the initial phase where a model learns general knowledge from vast amounts of data. The top techniques for this involve different objectives for the model to learn from. The most common is Causal Language Modeling (CLM), where the model predicts the next word in a sentence. Another is Masked Language Modeling (MLM), where the model fills in intentionally hidden words in a text. Other important methods include Denoising Objectives, where the model learns to restore a corrupted input to its original state, Contrastive Learning, where it learns to distinguish between similar and dissimilar pieces of data, and Next Sentence Prediction (NSP), where it determines if two sentences logically follow each other.

Fine-tuning Techniques: Fine-tuning is the process of adapting a general pre-trained model to a specific task using a smaller, specialized dataset. The most common approach is Supervised Fine-Tuning (SFT), where the model is trained on labeled examples of correct input-output pairs. A popular variant is Instruction Tuning, which focuses on training the model to better follow user commands. To make this process more efficient, Parameter-Efficient Fine-Tuning (PEFT) methods are used, with top techniques including LoRA (Low-Rank Adaptation), which only updates a small number of parameters, and its memory-optimized version, QLoRA. Another technique, Retrieval-Augmented Generation (RAG), enhances the model by connecting it to an external knowledge source during the fine-tuning or inference stage.

Alignment & Safety Techniques: Alignment is the process of ensuring an AI model's behavior aligns with human values and expectations, making it helpful and harmless. The most prominent technique is Reinforcement Learning from Human Feedback (RLHF), where a "reward model" trained on human preferences guides the AI's learning process, often using an algorithm like Proximal Policy Optimization (PPO) for stability. Simpler alternatives have emerged, such as Direct Preference Optimization (DPO), which bypasses the need for a separate reward model, and Kahneman-Tversky Optimization (KTO), which simplifies data collection further. To ensure safe deployment, Guardrails are implemented as a final safety layer to filter outputs and block harmful actions in real-time.

Enhancing AI Agent Capabilities

AI agents are systems that can perceive their environment and take autonomous actions to achieve goals. Their effectiveness is enhanced by robust reasoning frameworks.

Chain of Thought (CoT): This prompting technique encourages a model to explain its reasoning step-by-step before giving a final answer. This process of "thinking out loud" often leads to more accurate results on complex reasoning tasks.

Tree of Thoughts (ToT): Tree of Thoughts is an advanced reasoning framework where an agent explores multiple reasoning paths simultaneously, like branches on a tree. It allows the agent to self-evaluate different lines of thought and choose the most promising one to pursue, making it more effective at complex problem-solving.

ReAct (Reason and Act): ReAct is an agent framework that combines reasoning and acting in a loop. The agent first "thinks" about what to do, then takes an "action" using a tool, and uses the resulting observation to inform its next thought, making it highly effective at solving complex tasks.

Planning: This is an agent's ability to break down a high-level goal into a sequence of smaller, manageable sub-tasks. The agent then creates a plan to execute these steps in order, allowing it to handle complex, multi-step assignments.

Deep Research: Deep research refers to an agent's capability to autonomously explore a topic in-depth by iteratively searching for information, synthesizing findings, and identifying new questions. This allows the agent to build a comprehensive understanding of a subject far beyond a single search query.

Critique Model: A critique model is a specialized AI model trained to review, evaluate, and provide feedback on the output of another AI model. It acts as an automated critic, helping to identify errors, improve reasoning, and ensure the final output meets a desired quality standard.