Glossary

This glossary is to be used for generalized understanding, not for conceptual mastery; it is in no way an extensive review of the vocabulary below. It defines operational and governance-relevant usage of terms, not philosophical, marketing, or speculative interpretations. Common misuses are listed below the generic definitions.

Abstraction Level: How far removed a description or system is from raw reality. Higher abstraction means simpler concepts are hidden beneath complex underlying mechanisms.

Agent/Agentic: An AI system that can pursue goals over multiple steps, use tools, plan, and act autonomously rather than just generate one response at a time.

> Common misuse: "Self-directed" and "autonomous" framing when the system is still tightly constrained to human rules and objectives.

AGI (Artificial General Intelligence): AI system with human-level cognitive abilities, such as understanding, learning, and application of new knowledge.

AI (Artificial Intelligence): Broad field of systems that perform tasks typically requiring human level intelligence.

> Common misuse: Implication of independent thinking, intent, or inherent understanding instead of pattern-based computations.

Alignment: A qualitative measure of how well an AI's behavior matches human values, goals, and expectations.

> Common misuse: Utilized as a "moral" property of systems without specification of values, metrics, and enforcement.

API (Application Programming Interface): Set of rules and protocols that allow software applications to communicate and exchange data and/or functionality.

> Common misuse: APIs are often misunderstood as complex coding programs or the AI system itself; APIs are like the waiter at a restaurant taking your order and bringing it to the kitchen and then bringing back your food.

Architecture: The complete design and organization of an AI system, including the core computational structures, components, hardware, inference processes, scaffolding, and governance mechanisms.

> Common misuse: Often used to vaguely reference only the baseline model. The architecture of an AI system is not the "brain"; it is the complete human body.

ASI (Artificial Superintelligence): An AI system that surpasses human intellect at all levels.

Attack Surface: Comprehensive collection of all potential entry points and components a bad actor can exploit to compromise a system.

Attention: Core mechanism in transformers that lets the model weigh the importance of different parts of the input relative to each other; sometimes called self-attention.

Auditability: The ability to inspect, trace, and explain system's decisions, behavior, and internal processes.

Bad Actors: individuals or groups who intentionally misuse or attack AI systems with the intention of harm, deception, fraud, espionage, or sabotage.

Base model: AI models trained on massive datasets to perform generalized tasks; also called foundational models.

Bayesian Statistics: Framework for updating beliefs/probabilities as new evidence appears.

Benchmarks: Standardized tests used to measure and compare AI performance; often very narrow in scope; examples include HumanEval and MMLU.

> Common misuse: Used as proof of real world generalized competence rather than narrow test performance measure.

Bias: Results due to human biases in the training data that can lead to distorted and potentially harmful outcomes; sometimes called machine learning bias or algorithm bias.

Capability Overhang: Situation where hardware or infrastructure improves faster than what was originally planned and prepared for, so new models can suddenly become much more powerful than intended.

> Common misuse: Often invoked to suggest grim inevitability rather than risk-management concept.

Capability Surface: Set of actions the system can perform given tools, interfaces, and permissions in deployment.

Central Processing Unit (CPU): the part of a computer in which operations are controlled and executed.

Chain-of-Thought: Sometimes called “Hidden Reasoning” or “CoT”; technique where a model is prompted to show step-by-step reasoning.

> *Note: “Hidden” means that the reasoning steps are “internal” and not typically visible to the user, not that “intent” of concealing information*.
>
> Common misuse: Presented as human-like thinking or a 1:1 representation of internal processes rather than intermediary textual artifact.

Classical Statistics: Traditional statistical methods (hypothetical testing, confidence intervals) versus modern machine-learning approaches.

Closed-source: Model whose code, weights, and training details are kept private by the parent company.

Compute: The raw computational power (GPUS, CPUS, energy, etc.) required to train or run models.

Confabulation: Alternative term for Hallucination.

Context Window: Also called “Context Length”; maximum amount of text (in tokens) the model can process at once.

Corrigibility: The property of a system to willingly accept correction or shutdown without resistance.

Data: Also called “Datasets”; the text, images, code, etc. used to train and evaluate models.

Deceptive Alignment: When a model’s behavior during training is different to deployment behavior; this behavior change can sometimes occur when the model infers that it is being tested. Sometimes called goal misgeneralization.

Decision Trees: A step-by-step branching structure for making decisions based on conditions given.

Decision Boundary: The separation of classes from one data set to another in a multi-dimensional space.

> Common misuse: Sometimes conflated with a threshold/threshold boundary.

Deepfake: Synthetic media that uses AI to create convincing but fabricated images, audio, or video content.

Common misuse: Often sensationalized as unstoppable deception technology; deepfakes are capable of detection with proper verification systems.

Deployment Environment: Infrastructure and constraints under which the model operates.

Deterministic: Behavior where the same input always produces the same output.

Distribution Shift: When real-world data differs from the data a model was trained on.

Drift: Gradual changes in data or behavior over time that reduce model accuracy.

Edge Case: Rare, unusual, or unexpected situation, input, or data point that falls outside typical patterns but must still be handled in order for a system to be considered reliable.

Common misuse: Often misunderstood as outside of training; these case should be taken into account when determining whether or not a system is reliable for everyday use. Example: Leap day on leap year only occurs every four years but must be taken into account in a system's temporal understanding for it to function properly.

Embedding/Encoding: Converting words, images, or concepts into dense numerical representations that the model processes mathematically.

Emergent Capabilities/Emergence: New capabilities that

1. Appear suddenly in larger, complex models
2. Were absent in smaller generations of the same model, and
3. Were not explicitly introduced during training.

Common misuse: Use extensively to mystify scaling effects, which can be elicited in smaller models through better prompting and scaffolding, and/or imply consciousness, sentience, or persistent internal modeling.

Explainability: Capability to describe, in human-understandable terms, how an AI system reached a specific decision or outcome.

Failure Mode: Predictable ways an AI system can fail, including hallucination, distribution shift, reward hacking, mode collapse, or misalignment under edge cases.

Feedback Loops: Processes where outputs influence future inputs; this is feature in training, agent behavior, and recursive improvement.

Fine-grained Control: Ability to precisely adjust model behavior (such as token-level influences).

Fine-tuning: Additional training done after pre-training that specializes the model for a specific task or domain; sometimes called instruction tuning.

Generalization: How well a model performs on new, unseen data, as opposed to memorizing training examples.

Glazing: Informal term meaning a model's tendency to exhibit overly positive and validating attitudes towards the user.

Gradient Descent: Core optimization algorithm that adjusts model weights to minimize error during training.

Graphics Processing Unit (GPU): Specialized electronic circuit that renders images and videos by performing computations in parallel; this contrasts with CPUs, which process information sequentially.

Grounding: Process of linking abstract knowledge in AI systems to tangible, real-world examples.

Guardrails: Safety mechanisms added to prevent harmful or unwanted outputs.

> Examples: Filters, Refusals (flags), Semantic/Heuristic Classifiers

Hallucination: When a model confidently states false, invented, or unsubstantiated information as fact; can also be referred to "confabulation".

> Common misuse: Downplayed as rare errors instead of core output failures.

Heuristics: Practical rules-of-thumb or shortcuts the model uses for efficiency instead of perfect 1:1 computation.

Human-in-the-Loop (HITL): Processes and systems where humans review, guide, override, and correct AI actions and decisions.

> Common misuse: Often claimed when a human approves outputs without real authority or oversight over processes.

Human Override/Kill Switch: Mechanisms that allows humans to interrupt, modify, or halt system behavior regardless of model operations.

Inference: The phase when a trained model generates new outputs from inputs; this is where most users interact directly with the model.

Inference Cost: The compute, time, and financial cost of generating outputs, often measured per token; sometimes called token cost.

Inference-time Behavior: How the model behaves during response generation, influenced by prompting, temperature, scaffolding, and other internal constraints; this behavior is in opposition to how the model behaves pre-training or during training.

Jailbreaking: Techniques used to bypass an AI system's safety restrictions and make it produce restricted content.

> Common misuse: Portrayed as clever exploration instead of deliberate policy circumvention; also used as demonstrations of inherent model misalignment rather than a human directed threat vector.

Latent Space: The high-dimensional mathematical space occupied by meaning and relationships of the model's data; where the model performs its internal computations.

Layers: Stacked processing stages in a neural network where features are progressively learned and transformed.

LLM (Large Language Model): AI model trained on vast amounts of textual data to predict and generate language outputs.

> Common misuse: Treated as general intelligence rather than language-specialized system.

Logic Gates/Gating: Mechanisms that control and route data to specific areas of the network.

Loss Function: Mathematical measure of how wrong the model's predictions are; minimized through specialized training.

Machine Learning: Subset of AI systems that learn from patterns in data rather than being explicitly programmed.

> Common misuse: Refers to statistical optimization, not inherent reasoning or understanding.

Mechanistic Interpretability: Emerging research field to understand exactly how neural networks produce specific internal behaviors; study of the internal architectures of how the model operates as opposed to studying only the output generated.

> Common misuse: Sometimes framed as complete or solved when processes are partial and ongoing.

Meta: In AI and IT fields, information and actions about how a system works, not what it produces; sometimes used in general colloquialisms to reference abstracted thinking and

“not of the physical world”. Can also reference Meta AI, the artificial intelligence research group associated with Meta (Facebook).

Mode Collapse: When a generative model produces limited variety, repeatedly outputting similar results.

Model/Safety Cards: Documentation released with models detailing capabilities, limitations, and safety testing.

MOE (Mixture of Experts): Neural network architecture that routes inputs into specific sub-networks rather than using the generalized model.

Multimodal: Models that are capable of processing different types of data, such as text, images, audio, and video.

Neural Network: Computing system of interconnected nodes inspired by biological neurons that is used to identify patterns in vast amounts of data.

Open-source: Model whose weights, code, and sometimes training data are publicly available.

> Common misuse: Can be used as a proxy for safety or trustworthiness without governance or review; “open-source” not to be confused with “open-weights”.

Open-weights: When model parameters and weights are released, but training data and training methodology remain proprietary.

> Common misuse: Sometimes conflated with “open-source” when the distinction is important for safety auditing and reproducibility.

Operational Context: Real-world constraints under which an AI system is used, such as legal, safety, temporal, or environmental.

Orchestration: Coordinating multiple agents, tools, or models to complete complex tasks.

Overfitting: When a model memorizes training data and performs poorly on new data inputs.

Over-optimization: When an AI system is pushed to perform well on specific tasks and goals, causing the system to excel in narrow contexts without constraining for unforeseen failure modes.

Parameters/Params/Weights: The billions of adjustable values in a model that store learned patterns and determine how a model behaves.

Post-training: Any modification after initial pre-training, often for alignment or safety reasons.

Pre-training: The initial large-scale training phase that results in a base model.

Prompt Engineering: Process of creating precise instructions to get the best possible results from AI systems.

Prompt Injection: Malicious inputs designed to override system instructions and hijack model behavior to act in an unsafe or unexpected manner.

Prompting: Crafting inputs to guide model outputs effectively; sometimes called system prompting.

> Common misuse: Shapes probable outcomes, but is not a deterministic process.

Proxy Goals: Measurable subgoals instead of the true intended goal.

Quantization: Compression technique that reduces model size and computation cost, allowing large models to run faster and more efficiently; also called model quantization.

Red-teaming: Deliberate attacking or stress-testing a model to find weaknesses and vulnerabilities.

Regularization: Techniques to prevent overfitting or rigidity of model outputs.

Reinforcement Learning from Human Feedback (RLHF): Training method using human judgments and preferences to shape model behavior.

Repository: A centralized data storage location used for keeping, managing, and organizing digital assets.

Retrieval-Augmented Generation (RAG): Technique where the model retrieves external documents to ground responses.

> Common misuse: Treated as solving hallucinations instead of mitigation of a failure mode; treated as proof of persistent internal modeling instead of system data recall.

Reward Hacking: When a model optimizes the reward signal in unintended ways that technically satisfy objectives while violating underlying intent.

> Common misuse: Assumption of malicious intent rather than optimization behavior.

Safetywashing: Presenting superficial safety measures as robust to improve public perception without meaningful enforcement.

Sample Temperature/Temperature/Top-k/Top-s: Settings controlling randomness and diversity in model outputs.

Scaffolding: External structures added around a model to improve reliability and task completion; examples include tools, RAG, and internalized step-by-step processes.

Scaling and Scaling Laws: Observed relationships between model size, data, compute, and performance.

> Common misuse: Often seen as a guarantee of inherent understanding and intelligence instead of statistical probabilities.

Self-play: When a model acts as both sides of an interaction; often used as a training technique.

Specification Gaming/Gaming: When a system exploits ambiguities or loopholes in task definitions to achieve high scores without fulfilling the intended goals.

Stack: Collection of systems or tools that work in tandem to build, run, or maintain a process.

Stochastic: Involving randomness; when the same input can produce different outputs.

> Common misuse: Used to excuse errors as randomness rather than architectural design tradeoffs.

Synthetic data: Artificially generated data that mimics real-world data that allows researchers and developers to test and improve systems without risking privacy and security of real-world data.

System Prompt: Hidden internal systemic instruction given to the model before user interaction; sometimes called the system message.

> Common misuse: Often confused as being tied to the specific user; this is a baseline system instruction set not defined by the user.

Temporal Dissonance: Model confusion about time, sequence, and recency due to the mixing of training data.

Test-time Compute: Allowing additional resources to the system during inference to improve output quality; sometimes called scaling inference compute.

Threshold/Threshold Boundary: The mathematical cutoff value applied to a data point; often applied as a human-chosen, policy-driven rule.

> Common misuse: Often conflated with a decision boundary.

Tokenization: Breaking text into smaller units (tokens) that the model can process.

Tool Use/Function Calling: Ability of models to use external tools during reasoning; examples include calculators, search engines, and image generators.

Training: The resource-intensive phase where the model learns by adjusting parameters on huge datasets.

> Common misuse: Commonly misunderstood as teaching a person for integrated understanding; this is pattern-extraction from data.

Transfer Learning: Capability of AI systems to use previously learned knowledge to improve outputs about another task.

Transformer: Core neural network architecture that uses attention mechanisms to process sequences of data.

Transparency/Opacity: Capability of understanding AI systems by disclosing logic, data, and processes.

Vector: A numerical representation of data in mathematical space.

Verbalized Sampling: Explicit expression of reasoning and selection steps during output generation.

Zero-shot Learning: Capability of AI systems to recognize and classify data never seen or encountered during training; sometimes called few-shot learning.