Model Monster's AI Glossary

Generated February 25, 2026

This glossary is a shared vocabulary for discussing AI technology, deployment, and related governance, procurement, and policy topics. Definitions are intended to be primarily descriptive, with some pointers as to why certain terms matter in the legal context. However, these definitions are limited and how a term applies in a given matter depends on the facts, jurisdiction, and the governing agreement.

Context Tags:

[C] Common in contracting, procurement, product counseling, or vendor terms.

[R] Common in regulatory, compliance, or policy discussions.

[L] Common in litigation, legal doctrine, or adjudicatory contexts.

[T] Common in technical, engineering, or implementation discussions.

A

A2A

See also: Agent-to-Agent Protocol

Abuse [C, R] — Misuse of an AI system or service (e.g., policy-violating content, fraud, phishing, scraping, automated harassment) or abuse of the underlying service (e.g., credential stuffing, model extraction attempts). “Abuse” is commonly used in safety policies, AUP enforcement, and monitoring programs.

See also: Acceptable Use Policy; Content filtering; Model extraction; Rate limiting

Acceptable use — A general term for permitted use of a service. In many settings it refers to an Acceptable Use Policy (AUP) or similar contractual restrictions.

See also: Acceptable Use Policy; Terms of Service

Acceptable Use Policy (AUP) [C] — Rules (often incorporated into or referenced by Terms of Service) describing permitted and prohibited uses of an AI service and the provider’s enforcement options (e.g., warnings, suspension, termination). AUPs are commonly relevant to authorization questions, misuse response, and allocation of responsibility for prohibited uses. AUPs are frequently incorporated by reference into licenses.

See also: Content filtering; Safety policy; Terms of Service

Acceptance criteria [C] — Defined, testable conditions for determining whether a deliverable or service meets requirements (e.g., performance thresholds, evaluation results, latency, groundedness, safety tests). In AI agreements, acceptance criteria often tie to a specific intended use and evaluation methodology.

See also: Benchmark; Evaluation (evals); Intended use; SLA/SLO; Testing

Access control [C, R] — Technical and administrative mechanisms that limit who or what can access data, systems, and tools (e.g., authentication, authorization, role-based access control, least privilege). Access control is central to confidentiality, privacy, and security obligations.

See also: Authentication; Authorization; Connector; Data security; Least privilege

Accuracy — The proportion of predictions a model gets correct. Accuracy is context-dependent and often misleading without additional metrics: a model that predicts "no fraud" for every transaction achieves 99% accuracy if fraud occurs in only 1% of cases. Marketing claims citing accuracy should specify the dataset, task, and conditions; high accuracy on benchmarks may mask poor performance on important subgroups.

See also: Evaluation (evals); F1 score; Precision; Recall

Active learning — A machine learning approach where the model identifies which unlabeled examples would be most valuable to label, reducing annotation costs. Active learning strategies may involve human labelers viewing sensitive content, raising labor and content moderation considerations.

See also: Annotation; Human evaluation; Labeling

Adapter — A small set of parameters added to a frozen base model to customize behavior without full fine-tuning; includes techniques like LoRA. Adapters are often the deliverable in custom AI projects, so ownership, portability, and confidentiality terms are often stated explicitly. Frequently used as a way to provide organization-specific capabilities without creating full separate base models.

See also: Fine-tuning; LoRA (Low-Rank Adaptation); Parameter-efficient fine-tuning

Adversarial attack — An attempt to cause a model to produce incorrect or harmful outputs through crafted inputs. Adversarial robustness is relevant to security representations, product liability, possibly-infringing intellectual property outputs, and contractual performance standards. This can be an attack against the AI system as a whole or even just a prompt designed to elicit an unwanted result.

See also: Evasion attack; Jailbreak; Prompt injection; Red teaming

Adversarial example — A data instance purposefully perturbed to induce misclassification by a deployed model. Relevant to security testing, autonomous vehicle litigation, and content moderation failures.

See also: Adversarial attack; Data poisoning; Evaluation (evals); Robustness

Adversarial success — An AI system failure due to adversarial attack, where unwanted model output results in adverse effects like leakage of privileged data, violation of guardrails, expansion of privilege, or unwanted output. This defines the failure condition for many types of testing and is relevant to breach notification and incident response obligations.

See also: Data leakage; Jailbreak; Prompt injection

Adverse action [L, R] — A negative decision affecting an individual (commonly used in employment, credit, housing, insurance, and benefits contexts). AI-supported decisions that lead to adverse action may implicate notice, contestability, documentation, or anti-discrimination requirements depending on jurisdiction and sector.

See also: Algorithmic discrimination; Automated decision-making; Disparate impact

Agent / AI agent [C, R] — An embedded or dedicated software component that interacts with a person or system. An agent may or may not be "agentic" \- i.e., it may or may not include autonomous or semi-autonomous decision-making or tool-calling ability. The term “agent” is used loosely in the industry for any AI assistant.

See also: Agentic AI; Autonomy level; Multi-agent system; Tool calling (function calling)

Agent framework — Software infrastructure for building and deploying AI agents, often including tool integration, memory, and orchestration capabilities. Framework choice affects portability, vendor lock-in, and security posture; open source frameworks have different risk profiles than proprietary ones.

See also: Agentic AI; Orchestration; Tool calling (function calling)

Agentic AI [C, R] — An AI system that does one or both of the following: a) takes a larger request and breaks it down into smaller tasks for execution, and b) calls a tool (including another agent/system/model) and, based on the output, decides whether to provide a response, or continue (including changing the plan or calling another tool). Agentic AI systems require clear authorization boundaries, logging, human oversight mechanisms, and liability allocation for autonomous actions.

See also: Autonomy level; Excessive agency; Tool calling (function calling)

Agents Rule of Two — A security framework developed by Meta stating that AI agents should satisfy no more than two of the following three properties within a session: (A) processing untrustworthy inputs, (B) accessing sensitive systems or private data, and (C) changing state or communicating externally. Building on Simon Willison's Lethal Trifecta, the Rule of Two extends protection beyond data exfiltration to cover any state-changing action an agent might take, including examples like issuing refunds, modifying files, sending messages, or executing code. If a task requires all three properties, the agent should not operate autonomously and must include human-in-the-loop approval or equivalent supervision. The Rule of Two reflects the current consensus that prompt injection cannot be reliably detected or filtered, making architectural constraints the most practical defense for agentic AI systems.

See also: Agentic AI; Excessive agency; Human-in-the-loop; Least privilege; Lethal Trifecta; Prompt injection

Agent-to-Agent Protocol (A2A) [C] — An open source protocol enabling autonomous AI agents to discover, communicate, and collaborate with each other regardless of their underlying framework or vendor. A2A addresses the coordination layer: how agents delegate tasks, share capabilities, and track outcomes. Note: The A2A effort merged with IBM's Agent Communication Protocol (ACP).

See also: Agentic AI; Model Context Protocol; Multi-agent system; Orchestration

AI Bill of Materials (AIBOM) [C] — An inventory documenting the components, dependencies, and provenance of an AI system, analogous to an SBOM for software. An AIBOM typically covers the model(s) in an AI system and their versions, fine-tuning datasets, adapters, system prompts, guardrails, embedding models, vector databases, tools and plugins, and third-party services—the full CORE of the deployed system. AIBOMs support license compliance, supply chain security, incident response, and regulatory documentation requirements. They are increasingly requested in enterprise procurement and expected for high-risk AI systems.

See also: AI system; CORE; License compliance; Model card; Supply chain security

AI governance [C, R] — The policies, processes, roles, and controls for managing AI risk and ensuring responsible development and deployment. Regulators, customers, and boards increasingly expect AI governance programs, and support reasonable care arguments and compliance demonstrations.

See also: AI policy; Model registry; NIST AI RMF (AI Risk Management Framework); Risk assessment

AI model

See also: Model

AI policy [R] — Internal organizational policy or external governmental policy addressing development, deployment, procurement, and use of AI systems. Policies commonly cover permitted uses, data handling, human oversight, documentation, and incident response.

See also: Acceptable Use Policy; AI governance; Monitoring; Risk assessment

AI safety — Research and practices aimed at reducing harmful or unintended behavior of AI systems. In enterprise, procurement, and policy contexts, “AI safety” may refer to model training and post-training methods (e.g., RLHF), system-level guardrails, abuse monitoring, and evaluation programs tied to a stated threat model.

See also: Alignment; Guardrails; Red teaming; Safety policy

AI system [C, R] — The complete deployed software system, including all dataflow-affecting components, models, guardrails, control systems, accessible resources, allowed operations, and interfaces—not just the underlying model(s). An AI system's risk profile depends not only on model capabilities but on its CORE: the Components it comprises, the Operations it can perform, the Resources it can reach, and the Execution dataflow connecting them. Regulatory obligations under the EU AI Act and sector rules typically attach to the "AI system" as deployed, making it important to distinguish from "model" in contracts and governance.

See also: CORE; Deployer; High-risk AI system; Model; Provider

Algorithmic discrimination [L, R] — Differential treatment or outcomes produced by an AI system that may be associated with protected characteristics or other legally relevant categories. Depending on jurisdiction and use case, this concept can be relevant to civil rights, consumer protection, employment, housing, credit, and sector-specific obligations. Documentation and testing may be used to evaluate risk and compliance.

See also: Bias; Disparate impact; Fairness; High-risk AI system

Algorithmic Disgorgement (Model Deletion) [R] — A regulatory remedy requiring deletion of unlawfully obtained data and, in some cases, deletion of models or derived artifacts trained using that data (including weights, checkpoints, or fine-tunes). It may be sought where deleting the underlying dataset alone would not address downstream effects.

See also: Machine Unlearning; Prohibited AI practices; Training data

Alignment [R] — Ensuring AI system objectives and behaviors match human values, intentions, and instructions. Alignment encompasses both technical measures and governance, and affects safety representations and liability exposure.

See also: Constitutional AI; Safety policy; Value alignment

Annotation — The process of labeling data for training or evaluation; also called "tagging" or "labeling." Annotation involves human labor (often outsourced), content exposure, and quality control issues; IP questions arise for labeled datasets.

See also: Ground truth; Labeling; Training data

Anonymization [R] — A form of data processing intended to make data no longer identifiable to any individual, typically requiring that re-identification is not reasonably likely given available means. Legal definitions and thresholds vary by jurisdiction and context.

See also: De-identification; Personal data; Pseudonymization

Answer relevance — A retrieval-augmented generation (RAG) evaluation measure assessing how well an answer addresses the user's question, often compared to a reference or judged by humans/LLMs. It is frequently used alongside context relevance and groundedness.

See also: Context relevance; Evaluation (evals); Groundedness

Application Programming Interface (API) [C] — A structured interface allowing software to communicate with other software programmatically, as distinguished from a human-facing interface like a website or app. In the AI context, API access means integrating AI capabilities directly into applications, workflows, or products through code rather than through a chat interface. API terms typically differ significantly from consumer terms: they often permit broader commercial use and integration but impose rate limits, usage-based pricing, data handling obligations, and restrictions on downstream redistribution. Key contractual issues include whether outputs can be used to train competing models, what data is logged and retained, SLA commitments, and how usage is metered and billed.

See also: Endpoint; Rate limiting; Service Level Agreement / Service Level Objective; Usage data / telemetry; Zero Data Retention

Approval workflow [C] — A process requiring human authorization before certain actions or outputs are executed or released. Approval workflows are key controls for high-risk agentic AI systems.

See also: Autonomy level; Change control; Human-in-the-loop

Architecture — The high-level structure of a model or AI system and how components interact (e.g., model \+ retrieval \+ tools \+ guardrails \+ monitoring). Architecture choices influence performance, security, privacy, and auditability. One way to identify the parts of an architecture is by using the mnemonic CORE \- Components, Operations, Resources, and Execution.

See also: AI system; CORE; Guardrails; Tool calling (function calling)

Artificial General Intelligence (AGI) — A non-standard term used to describe hypothetical AI with broad, human-level capability across many domains. The term is used inconsistently in technical and marketing contexts; many current systems are better described as foundation models or general-purpose AI rather than "general intelligence." In diligence and policy discussions, "AGI" sometimes signals discussion of frontier capability thresholds and risk controls.

See also: Capabilities; Foundation model; Narrow AI

Attention mechanism [T] — The method by which transformer models determine which parts of an input matter for producing each part of an output. When generating the next word, the model assigns weights to every previous token, attending more to relevant context and less to irrelevant text. Attention is not comprehension; the model is computing statistical relevance, not understanding meaning. The attention patterns can sometimes be examined to understand why a model produced particular outputs, though this interpretability has limits.

See also: Context window; Self-attention; Transformer

Audit [C, R] — Independent examination of AI systems, processes, or controls for compliance, performance, or risk assessment. Audit rights are common in enterprise contracts. Third-party AI audits are emerging as a governance and compliance tool.

See also: AI governance; Conformity assessment; Evaluation (evals); SOC 2

Audit log — A record of system events that supports tracing actions, changes, and access (e.g., who accessed data, what tools were called, what model version ran). Audit logs are commonly used for security investigations, compliance, and dispute resolution.

See also: Access control; Change control; Logging; Monitoring

AUP

See also: Acceptable Use Policy

Authentication — Verification of user or system identity before granting access to AI services or data. Authentication controls are baseline security requirements; failures can create breach liability and confidentiality exposure.

See also: Access control; Security

Authorization — The process of determining what an authenticated user, service, or agent is allowed to do (permissions), often implemented via roles, policies, or scopes. Authorization differs from authentication (identity verification).

See also: Access control; Authentication; Least privilege

Autoencoder [T] — A neural network that learns compressed representations by encoding inputs and then reconstructing them. Used in anomaly detection, data compression, and generative models; relevant when understanding technical architecture of certain AI systems.

See also: Decoder; Encoder; Latent space

Automated decision-making (ADM) [R] — Decisions made by AI systems with limited or no human involvement, often subject to regulatory requirements. GDPR Article 22, various US state laws, and sector regulations impose requirements on ADM including rights to explanation and human review.

See also: Explainability; Human-in-the-loop; Right to explanation

Autonomy level [C, R] — The degree of independence an AI system has in making and executing decisions without human oversight. Autonomy level is central to risk assessment, liability allocation, and regulatory compliance for agentic systems.

See also: Agentic AI; Approval workflow; Human-in-the-loop

Availability [C] — The percentage of time a system is operational and accessible, often defined in SLAs. Availability commitments for AI-dependent systems often account for model-specific issues (capacity, rate limits, provider outages).

See also: Business continuity; Service Level Agreement / Service Level Objective; Uptime

B

Backdoor (model backdoor / trojan) — A hidden behavior in a model or system that activates under specific triggers (e.g., particular phrases, patterns, or inputs), causing unintended outputs or actions. Backdoors are discussed in model security, supply chain risk, and red teaming contexts.

See also: Adversarial attack; Data poisoning; Model supply chain; Red teaming

Backpropagation — The algorithm used to train neural networks by computing how much each weight contributed to prediction errors, then adjusting weights to reduce those errors. Backpropagation is how models "learn" from training data; errors propagate backward through the network, and weights are updated accordingly.

See also: Gradient descent; Training; Weights

Base model [T] — A pre-trained model before any fine-tuning or customization; the starting point for downstream adaptations.

See also: Fine-tuning; Foundation model; Pre-training

Batch size — The number of training examples processed together before updating model weights; a hyperparameter affecting training dynamics. Affects training resource requirements and costs.

See also: Epoch; Hyperparameter; Training

Benchmark [C, T] — A standardized test for measuring model performance on specific tasks, enabling comparison across models. Benchmark claims in marketing often specify which benchmark, version, and conditions; benchmarks may not reflect real-world performance.

See also: Accuracy; Evaluation (evals)

Bias [L, R] — Systematic differences in model behavior or error rates that correlate with particular features, groups, or contexts. Bias is not limited to interactions with humans; it reflects the degree to which the distribution of features in training data matches the distribution in production. A model trained primarily on certain populations, document types, or scenarios will perform differently on others. In regulated contexts, bias testing and mitigation records are used to assess compliance posture.

See also: Algorithmic discrimination; Disparate impact; Fairness

Biometric data [R] — Data derived from physical or behavioral characteristics used for identification (face, voice, fingerprint, gait). Biometric data triggers heightened obligations under BIPA, GDPR, state privacy laws, and the EU AI Act; AI systems processing biometrics require special controls.

See also: Multimodal model; Personal data; Privacy

Black box — A system whose internal decision process is difficult to interpret or explain in a human-understandable way. The term is used in technical, governance, and legal contexts when evaluating transparency, accountability, and auditability.

See also: Explainability; Model card; System card; XAI

Business continuity [C] — Plans and capabilities for maintaining operations during disruptions, including AI system failures or provider outages. AI-dependent workflows need continuity planning; contracts commonly address provider failures, model deprecation, and data portability.

See also: Availability; Portability; Vendor lock-in

C

Capabilities — The tasks and performance characteristics a model or system can reliably support (e.g., summarization, coding, extraction, tool use). “Capabilities” is often tied to evaluation results and can be discussed in procurement, marketing, and policy settings.

See also: Benchmark; Evaluation (evals); Intended use; Model capability

Chain of Thought (CoT) [T] — A prompting technique that encourages models to show intermediate reasoning steps, often improving accuracy on complex tasks. CoT reasoning may provide some transparency into model "thinking" but is not a substitute for true explainability; reasoning traces may be fabricated.

See also: Explainability; Prompting; Reasoning model

Change control [C] — Processes governing modifications to AI systems, including model updates, prompt changes, and configuration adjustments. Change control is essential for regulated deployments; contracts commonly specify notice, approval, testing, and rollback requirements.

See also: Model drift; Model update; Version pinning

Chat model [T] — An LLM configured or post-trained for conversational interaction (e.g., instruction following, dialogue safety behaviors). Chat models are often accessed through chat-completion interfaces and may differ from base models in behavior and safety characteristics. ChatGPT was the first widely known chat model and remains the best-known example.

See also: Instruction tuning; Large Language Model; System prompt

Checkpoint — A saved state of model weights during training or fine-tuning, used to resume training or to preserve intermediate versions. In training agreements, checkpoints may be deliverables; agreements often specify ownership, retention, access controls, and permitted reuse.

See also: Model artifact; Training; Weights

Chemical, Biological, Radiological, Nuclear (CBRN) — Categories of weapons or hazardous materials; AI systems' potential to assist with CBRN threats is a key safety concern. CBRN-related content is typically prohibited in AUPs; frontier AI safety evaluations specifically test for CBRN assistance capabilities.

See also: Dual-use; Prohibited AI practices; Safety evaluation

Child Sexual Abuse Material (CSAM) [R] — Illegal content depicting sexual abuse of minors. Many AI services implement controls to detect, block, and report CSAM due to legal obligations and platform policy requirements; specific duties, reporting pathways, and retention requirements vary by jurisdiction and service design.

See also: Content filtering; Prohibited AI practices; Safety policy

Chunking — Dividing documents into smaller segments for processing within context window limits or for retrieval purposes. Chunking strategies affect retrieval accuracy and completeness; relevant when assessing whether AI systems properly considered full documents.

See also: Context window; Truncation

Citation — A reference to a source used to support a statement or output (e.g., a retrieved document chunk in RAG, or a legal citation to authority). In AI systems, “citations” may be generated automatically and can be incorrect or incomplete unless the system is designed to capture provenance.

See also: Grounding; Hallucination; Source attribution

Classification — The task of assigning inputs to predefined categories (e.g., spam detection, sentiment analysis, content moderation). Classification errors have different consequences depending on the application; false positives and negatives have different risk profiles.

See also: Accuracy; Discriminative model

Cloud provider [C] — A vendor providing cloud infrastructure and managed services (compute, storage, networking) used to train or operate AI systems. Cloud provider terms commonly affect data residency, security controls, incident response, and subprocessors.

See also: Data residency; Hosting; Security addendum; Subprocessor

Clustering — Grouping similar items together based on their features without predefined labels. Clustering can produce de facto sensitive inferences (grouping by health, demographics) even without explicit attributes.

See also: Embedding; Privacy; Unsupervised learning

CNN

See also: Convolutional Neural Network

Coalition for Content Provenance and Authenticity (C2PA) — A technical standard for embedding provenance information in digital content, including AI-generated content. C2PA and similar standards support content authenticity verification; increasingly relevant for deepfake detection and evidence authentication.

See also: Content provenance; Deepfake; Watermarking

Components — In the CORE framework, the functional elements that comprise an AI system: models, adapters, guardrails, databases, APIs, connectors, human review steps, and other nodes through which data flows. Each component has properties relevant to governance: its provider or origin, the operations it performs, the resources it accesses, and how it transforms the data flowing through it. A list of components is the minimum information needed for an AIBOM.

See also: Adapter; AI system; CORE; Guardrails; Model

Compute [C, R] — The computational resources (processing power, memory, storage) required to train and run AI systems; also a regulatory concept. Compute thresholds trigger reporting requirements under various Executive Orders, state laws, and the EU AI Act. Compute access is a key factor in AI capabilities.

See also: Export controls; Inference; Training

Computer vision [T] — A field of AI focused on interpreting and generating information from images and video (e.g., object detection, segmentation, captioning). Many modern systems use multimodal models that combine vision and language capabilities.

See also: CNN; Multimodal; Vision-language model

Confidential computing — Hardware-supported techniques protecting data while in use (in memory) using secure enclaves. Confidential computing may support stronger security representations for processing sensitive data in cloud environments.

See also: Encryption; Security; Trusted execution environment

Confidential information [C] — Information protected under contract or law from unauthorized use or disclosure; in AI contexts often includes prompts, outputs, logs, and fine-tuning artifacts.

See also: Logging; Trade secret; Usage data / telemetry

Confidentiality — Protection of non-public information from unauthorized disclosure. In AI contracting, confidentiality terms often address prompts, outputs, logs, and resources available to the AI system such as connected enterprise data.

See also: Access control; Confidential information; Logging

Conformity assessment [R] — A structured assessment process (often regulatory) to verify that a system meets specified requirements before deployment. EU AI Act and some sector regulations require conformity assessments for higher-risk AI; determine who performs them and what evidence is retained.

See also: EU AI Act; High-risk AI system; Technical documentation

Connector — An integration that pulls content from enterprise systems (e.g., SharePoint, Google Drive, Slack) into an AI system for retrieval or context. Connectors expand the data-access surface area; permissions, logging, and retention practices are commonly evaluated to reduce privilege, confidentiality, and privacy risk.

See also: Access control; CORE; Knowledge base; Operations; Resources

Consent [L, R] — Agreement by an individual to a specified use of their data, likeness, or content. In AI contexts, consent may be relevant to data collection, biometric processing, voice cloning, and digital replicas; requirements vary by jurisdiction and sector.

See also: Biometric data; Digital Replica; Personal data; Right of Publicity

Constitutional AI — An alignment approach using a set of principles ("constitution") to guide model behavior, often using AI-generated feedback. Constitutional policies can be relevant to content and safety representations; request documentation for high-stakes use cases.

See also: Alignment; Safety policy

Content filtering [C] — Automated detection and blocking or transformation of disallowed content. Filtering affects safety claims, regulatory compliance, intellectual property infringement, and AUP enforcement; raises false positive/negative issues.

See also: Guardrails; Moderation; Safety policy

Content provenance [R] — Information describing the origin and transformation history of content, including whether AI was involved in creation. Provenance supports authenticity, IP compliance, and consumer transparency; increasingly relevant for evidence authentication and misinformation disputes.

See also: Deepfake; Metadata; Watermarking

Context overload attack — An adversarial technique overloading the prompt with excessive tokens to predispose models to a vulnerable state. A specific prompt injection variant; relevant to security testing and incident analysis.

See also: Adversarial attack; Context window; Prompt injection

Context relevance — A RAG quality metric measuring whether retrieved context contains information pertinent to the user's query. Poor context relevance can cause unreliable outputs; relevant when evaluating RAG system performance claims.

See also: Answer relevance; Groundedness

Context window [C, T] — The maximum number of tokens a model can consider at once, encompassing both the input and the output being generated. Context windows have expanded dramatically–from about 4,000 tokens in early GPT-4 to over 1 million tokens in some current models–but limits still matter. When input exceeds the context window, content is truncated, often without notification to the user. Context window size is distinct from how well a model uses that context; performance often degrades on information buried in the middle or nearer to the end of long inputs.

See also: Token; Truncation

Continual learning — A learning paradigm where AI systems incrementally learn from new data while preserving prior knowledge (avoiding "catastrophic forgetting"). Continual learning systems may evolve in ways that affect prior representations about behavior; governance often addresses ongoing changes.

See also: Model drift; Model update; Training

Continuous learning

See also: Continual learning

Contract [C] — A legally binding agreement defining rights, obligations, and remedies between parties. In AI contexts, contracts often define scope of use, data rights, confidentiality, IP terms, warranties, indemnities, SLAs, and change control.

See also: Indemnity; Limitation of liability; SLA/SLO; Terms of Service

Controller / processor [C, R] — Privacy law roles: controller determines purposes/means of processing; processor processes on behalf of controller. AI vendors often characterize themselves as processors; customers may require controls on model training and subprocessing consistent with that role.

See also: DPA; Personal data; Subprocessor

Controls — Technical or organizational measures used to achieve defined objectives (e.g., security controls, privacy controls, safety controls). In governance and audits, “controls” are often documented, tested, and monitored. In contrast to regular GRC systems, AI controls need to be implemented with technical measures, frequently as components external to the model.

See also: Audit; Monitoring; Privacy-enhancing technology; Security

Convolutional Neural Network (CNN) [T] — A neural network architecture designed for processing grid-like data (images), using convolutional layers to extract features. CNNs power most computer vision applications including facial recognition and autonomous vehicles.

See also: Computer vision; Deep learning; Neural network

Copyleft license [C] — Open source licenses that may require distributing source code or licensing downstream when distribution triggers occur (e.g., GPL). Copyleft obligations can create compliance risk when AI systems distribute software or embed licensed components.

See also: License compatibility; Open source

Copyright [L] — A body of law protecting original works of authorship fixed in a tangible medium, granting exclusive rights (e.g., reproduction, distribution, derivative works) subject to limitations and exceptions. In AI discussions, copyright commonly arises with training data provenance, output ownership, and infringement/fair use analysis.

See also: Copyright infringement (AI context); Fair use; Output; Training data

Copyright infringement (AI context) [L] — Claims that AI outputs reproduce copyrighted material or that AI training constituted infringement. Central issue in pending litigation; training data provenance, output filtering, and indemnification are key contractual topics.

See also: Fair use; IP indemnity; Substantial similarity; Training data

CORE [C] — A framework and mnemonic device for analyzing and documenting AI systems by mapping their Components, Operations, Resources, and Execution dataflow. CORE represents AI systems as directed graphs where data flows through connected elements, enabling policy evaluation, compliance tracking, and risk assessment over an entire AI system.

See also: AI governance; AI system; Components; Execution; Operations; Resources

Cosine similarity — A metric comparing embeddings by measuring the angle between vectors; higher similarity implies closer semantic meaning. Relevant when understanding RAG retrieval mechanics; similarity thresholds affect retrieval defensibility.

See also: Embedding; Semantic search; Vector database

CoT [T]

See also: Chain of Thought (CoT)

Cross-border data transfer [C, R] — Moving personal data across national borders, triggering transfer mechanisms and restrictions. AI vendors may route prompts/logs across regions; DPAs and data residency terms often match actual architecture.

See also: Data residency; DPA; Privacy

CVE [C]

See also: Vulnerability / CVE

Cyber Resilience Act (CRA) [R] — EU Regulation 2024/2847 establishing mandatory cybersecurity requirements for "products with digital elements" (hardware and software connected to devices or networks) sold in the EU market. The CRA entered into force December 2024, with full applicability by December 2027\. It requires manufacturers to ensure products are secure by design, maintain vulnerability management throughout the product lifecycle, provide security updates, and report actively exploited vulnerabilities. Products are classified by risk level (critical, important, or default), with higher-risk products requiring third-party conformity assessment. The CRA applies to most software including AI systems and their components; it intersects with the EU AI Act (which addresses AI-specific risks) and requires SBOM-like documentation of components. Open source software developed outside commercial activity is generally exempt, though commercial products incorporating open source remain in scope.

See also: EU AI Act; Security; Supply chain security; Vulnerability / CVE

D

Data annotation

See also: Annotation

Data augmentation — Techniques that expand or vary training data to improve model generalization (e.g., transformations, paraphrases, synthetic examples). Augmentation can affect performance and bias characteristics depending on how it is applied.

See also: Bias; Synthetic data; Training data

Datacenter — A facility housing computing infrastructure (servers, storage, networking) used for training and operating AI systems. Datacenter location and subcontracting can matter for data residency, security, and resilience.

See also: Business continuity; Cloud provider; Data residency

Data classification [C, R] — An organizational scheme for labeling resource data by sensitivity and handling requirements (e.g., public, internal, confidential, regulated). Data classification commonly drives AI policy decisions about prompts, connectors, logging, and vendor use.

See also: Confidential information; Connector; Data governance; Logging; Resources

Data controller

See also: Controller / processor

Data drift — Changes over time in the distribution or characteristics of input data (or user behavior) that can degrade model performance. Data drift is often monitored alongside model drift (changes in the model itself).

See also: Evaluation (evals); Model drift; Monitoring

Data governance [C, R] — Organizational processes and controls for managing data quality, access, lineage, retention, and compliance. Data governance often sets the baseline for AI policy, connector use, logging, and training restrictions.

See also: Access control; Data classification; Retention; Training data

Data leakage [C] — Unintended disclosure of training data, prompts, or other sensitive information through model outputs. Data leakage can breach confidentiality, privacy, and security obligations; memorization and extraction attacks are key concerns.

See also: Membership inference; Memorization; Model inversion

Data Loss Prevention (DLP) — Technologies and policies preventing unauthorized data transmission or exposure. DLP tools can help prevent sensitive data from being sent to AI systems; relevant for shadow AI governance.

See also: Confidential information; Security; Shadow AI

Data minimization [R] — A privacy principle requiring collection and retention of only data necessary for specified purposes. Data minimization applies to AI training, inference logging, and improvement uses; conflicts with desires for comprehensive data may arise.

See also: Privacy by design; Purpose limitation

Data poisoning — An adversarial attack inserting or modifying training data to compromise model behavior. Data poisoning attacks can undermine model integrity; supply chain security and data provenance controls are relevant defenses.

See also: Adversarial attack; Supply chain security; Training data

Data privacy attack — Attacks designed to gain access to sensitive information in training data, including membership inference and model inversion. Privacy attacks demonstrate that training data can sometimes be extracted; relevant to privacy representations and security controls.

See also: Data leakage; Membership inference; Model inversion

Data Processing Addendum (DPA) [C] — A contract addendum addressing personal data processing requirements, typically required by privacy regulations. DPAs often address AI-specific issues: training uses, inference logging, subprocessors, model improvement, and cross-border transfers.

See also: Controller / processor; Personal data; Subprocessor

Data processor

See also: Controller / processor

Data Protection Impact Assessment (DPIA) [R, C] — An assessment required under GDPR Article 35 before processing likely to result in high risk to individuals' rights and freedoms, including systematic profiling, large-scale processing of sensitive data, and systematic monitoring. DPIAs must describe the processing, assess necessity and proportionality, identify risks, and specify mitigations. AI systems frequently trigger DPIA requirements due to automated decision-making, profiling, and processing at scale. Unlike FRIAs (which address broader fundamental rights), DPIAs focus specifically on data protection risks.

See also: Automated decision-making; Personal data; Privacy

Data Protection Officer (role) (DPO) [R] — A designated individual responsible for overseeing GDPR compliance, advising on data protection obligations, and serving as the point of contact for supervisory authorities and data subjects. Not to be confused with Direct Preference Optimization (DPO) (training).

See also: Controller / processor; DPO; Privacy

Data provenance — Documentation of data origin, collection methods, transformations, and chain of custody. Provenance is essential for IP compliance, bias assessment, and regulatory documentation; request provenance information for training data.

See also: Content provenance; Dataset documentation; Training data

Data reconstruction — A privacy attack designed to reconstruct sensitive information from training data through model queries. Reconstruction attacks demonstrate privacy risks from training; relevant to data sensitivity assessments.

See also: Data privacy attack; Memorization; Model inversion

Data residency [C, R] — Requirements that data be stored or processed in specific geographic locations. Data residency requirements may constrain AI architecture choices; verify that vendor infrastructure meets residency requirements.

See also: Cross-border data transfer; DPA; Subprocessor

Data retention [C] — Policies governing how long data is kept before deletion. Retention of prompts, outputs, and logs creates ongoing risk; align retention with purpose limitation and deletion rights.

See also: Logging; Privacy; Records retention

Data security [C, R] — Safeguards protecting data confidentiality, integrity, and availability, including access controls, encryption, logging, incident response, and secure development practices. In AI systems, data security applies to prompts, outputs, logs, embeddings, and connected enterprise data.

See also: Access control; Encryption; Security; Security addendum

Data segregation [C] — Separation of customer data or environments to prevent unauthorized access or cross-tenant leakage (logical or physical). Segregation is relevant to multi-tenant AI services, RAG indexes, and logging systems.

See also: Access control; Confidentiality; Data residency; Multi-tenant

Dataset documentation [T] — Structured information about a dataset's contents, collection, limitations, and intended uses (e.g., datasheets, data cards). Dataset documentation supports IP diligence, bias assessment, and regulatory compliance; request it for training datasets.

See also: Data provenance; Model card; Training data

Data subject [R] — An identified or identifiable individual whose personal data is processed. In privacy frameworks (e.g., GDPR and state privacy laws), data subjects may have rights (access, deletion, correction, objection, portability, etc.), and organizations typically implement processes to respond to those rights.

See also: Personal data; Privacy

Decisioning [C] — Using AI outputs to make or inform consequential decisions about individuals or transactions. Decisioning triggers regulatory obligations, liability exposure, and fairness requirements; distinguish advisory outputs from decisioning.

See also: Adverse action; Automated decision-making; High-risk AI system

Decoder [T] — A neural network component that reconstructs outputs from compressed representations. Decoders are the generative component in many AI architectures.

See also: Autoencoder; Encoder; Transformer

Decoder-only model [T] — A transformer model using only the decoder, suited for text generation tasks (e.g., GPT family). Most modern LLMs are decoder-only architectures.

See also: Encoder-only model; LLM; Transformer

Deepfake [L, R] — Synthetic media created using AI to depict people saying or doing things they did not actually say or do. Deepfakes raise defamation, fraud, election, NCII, and evidence authentication issues; detection and provenance tools are evolving.

See also: Content provenance; Generative AI (GenAI); NCII; Synthetic media

Deep learning — Machine learning using neural networks with multiple layers, enabling detection of complex patterns in data. The term is often used interchangeably with "AI" in business contexts, though technically it refers to a specific architectural approach. The "deep" in deep learning refers to the number of layers between input and output, not to any quality of understanding.

See also: Machine learning; Neural network; Training

De-identification [R] — Techniques that remove or obscure identifiers to reduce the ability to link data to a specific individual. De-identification is commonly evaluated based on the risk of re-identification given available auxiliary data, threat models, and technical safeguards; requirements and standards vary by law and context.

See also: Anonymization; Personal data; Pseudonymization

Deletion — Removal of data from systems and backups according to defined policies and technical processes. In privacy and AI contexts, deletion can relate to data subject rights, retention policies, and requests to remove training data or derived artifacts.

See also: Algorithmic Disgorgement (Model Deletion); Data subject; Machine Unlearning; Retention

Deliverables [C] — Items a party is obligated to provide under an agreement (e.g., fine-tuned model, adapter, documentation, evaluation results, training logs, or a deployed service). In AI projects, deliverables are often defined to clarify ownership, acceptance, and maintenance responsibilities.

See also: Acceptance criteria; Documentation; Model artifact

Deployer [R] — A party that deploys an AI system for use, as distinguished from the provider/developer. EU AI Act and other frameworks allocate different obligations to providers vs. deployers; determine your role and resulting duties.

See also: AI system; EU AI Act; Provider

Deterministic — A system property where the same input always produces the same output. Traditional software is deterministic by design. AI systems can in theory be configured for deterministic behavior, though hardware and infrastructure variations may still introduce variability. Deterministic operation supports reproducibility, testing, audit, and regulatory compliance, but may reduce output quality or diversity compared to default settings.

See also: Non-deterministic; Reproducibility; Sampling; Temperature

Developer prompt [T] — Instructions provided by an application or developer that guide model behavior for a specific product or workflow (e.g., formatting, tool use rules, safety constraints). Developer prompts typically have lower priority than system prompts but higher priority than user prompts.

See also: Instruction hierarchy (prompt precedence); System prompt; User prompt

Development practices — The processes and standards used to build and maintain software and AI systems (e.g., secure coding, testing, review, documentation, incident response). These practices are often described in security questionnaires, audits, or contractual representations.

See also: Documentation; Secure development lifecycle; Testing

DevOps — Practices combining software development and operations to improve deployment frequency, reliability, and monitoring (e.g., CI/CD, infrastructure as code). In AI settings, DevOps is often paired with MLOps for model lifecycle management.

See also: Change control; MLOps; Monitoring

Differential privacy [R] — A privacy technique that adds controlled statistical noise to limit what can be inferred about any individual record from an aggregate output. Differential privacy is discussed in privacy engineering, dataset release, and sometimes in training/evaluation contexts.

See also: De-identification; Personal data; Privacy-enhancing technology

Diffusion model [T] — A generative AI architecture that creates outputs by iteratively denoising random noise, commonly used for image generation. Diffusion models power DALL-E, Midjourney, Stable Diffusion; they raise distinctive copyright and provenance questions for images.

See also: Generative AI (GenAI); Image generation; Text-to-image

Digital Replica [L, R] — A term used in right-of-publicity and synthetic media discussions to describe a computer-generated representation of a person’s image, voice, or likeness. Applicable requirements and remedies vary by jurisdiction and may depend on consent, context, and whether the replica is used for commercial or deceptive purposes.

See also: Deepfake; Right of Publicity; Synthetic media

Direct Preference Optimization (training) (DPO) [T] — A training technique that aligns models using preference data without requiring a separate reward model. DPO is an alternative to RLHF for alignment training. Not to be confused with Data Protection Officer (DPO) (role).

See also: Alignment; Training

Disclosure — Communication of information to a user, counterparty, regulator, or the public (e.g., about AI use, limitations, data practices, or incidents). Disclosure duties can arise from contracts, consumer protection rules, sector regulations, or internal governance.

See also: Documentation; Notice; Transparency

Discriminative model — A model that classifies inputs or distinguishes between categories rather than generating new content. Discriminative models (classifiers, detectors) have different risk profiles than generative models; errors are often binary.

See also: Classification; Generative AI (GenAI)

Disparate impact [L, R] — Facially neutral practices that disproportionately affect protected groups. Disparate impact analysis applies to AI systems used in employment, credit, housing; testing often assess group-level outcomes.

See also: Algorithmic discrimination; Bias; Fairness

Distillation [T] — Training a smaller "student" model to mimic a larger "teacher" model's behavior by learning from the teacher's outputs rather than the original training data. Distillation does not require access to the teacher model's weights, only the ability to query it and observe its outputs. This creates trade secret and competitive concerns: even without sharing weights, unguarded API access may allow third parties to replicate proprietary model capabilities. Distillation can also transfer copyrighted expression if the teacher's outputs are used as training data. Many model licenses and API terms of service explicitly prohibit using outputs to train other models.

See also: AUP; Knowledge transfer; Model compression; Model extraction; Trade secret; Training; Weights

Distributional robustness — A model's ability to perform equitably across the range of possible inputs, including rare or "long-tail" cases. Distributional robustness affects fairness and reliability; models may fail on underrepresented populations or edge cases.

See also: Evaluation (evals); Fairness; Robustness

Documentation — Written materials describing a model, system, dataset, or process (e.g., model cards, system cards, technical specs, policies, runbooks). Documentation is commonly used for governance, audits, contracting, and incident response.

See also: Audit; Dataset documentation; Model card; System card

DPA

See also: Data Processing Addendum

DPO [T]

See also: Data Protection Officer (role); Direct Preference Optimization (training)

Dual-use [R] — Technology that has both legitimate and potentially harmful applications. Dual-use considerations affect export controls, safety evaluations, and responsible deployment decisions for frontier AI.

See also: Dual-use foundation model; Export controls

Dual-use foundation model [R, T] — A foundation model that exhibits or could be modified to exhibit high levels of performance at tasks posing serious security risks. Definition from Executive Order 14110; triggers reporting requirements and enhanced safety obligations.

See also: Executive Orders on AI; Foundation model

E

Edge deployment — Running AI models on local devices rather than cloud servers. Edge deployment affects data residency, latency, security, and update mechanisms; may be preferred for privacy-sensitive applications.

See also: Data residency; Latency; On-prem deployment; Small Language Model

E-discovery [L] — The process of identifying, preserving, collecting, and producing electronically stored information (ESI) in litigation or investigations. AI system logs, prompts, outputs, tool calls, and model/version records can be relevant ESI depending on the matter.

See also: Audit log; Litigation hold; Logging; Retention

Embedding [C, T] — A list of numbers (vector) representing the meaning of text, images, or other content in a form that enables mathematical comparison. Two pieces of text with similar meanings will have similar embeddings, allowing systems to find semantically related content even without shared keywords. Embeddings power RAG retrieval: when a user asks a question, the system converts it to an embedding and finds stored documents with nearby embeddings.

See also: Semantic search; Vector; Vector database

Embedding model [T] — A model specifically designed to generate embeddings for retrieval and similarity tasks. Embedding model selection affects retrieval quality and privacy risk (what information is encoded).

See also: Embedding; Vector database

Emergent capabilities — Unexpected abilities that appear in larger models without explicit training for those tasks. Emergent capabilities complicate capability assessment and safety testing; models may have abilities not anticipated at launch.

See also: AI safety; Evaluation (evals); Scaling law

Encoder [T] — A neural network component that compresses inputs into lower-dimensional representations. These compressed representations are used for retrieval and other tasks.

See also: Decoder; Embedding; Transformer

Encoder-only model [T] — A transformer model using only the encoder, suited for classification and extraction tasks. Encoder-only models are used for classification, not generation.

See also: Classification; Decoder-only model; Transformer

Encryption [C] — Cryptographic protection of data at rest and in transit. Encryption is baseline security for AI systems; clarify whether prompts, outputs, and logs are encrypted.

See also: Confidential computing; Data security; Security

End-of-life / deprecation (EOL) — The point at which a vendor or developer stops supporting a model, API, or product version (e.g., no updates, limited availability, retirement). Deprecation affects continuity, change control, and portability planning. Despite broad improvement in model capabilities, the deprecation of a model can change risks in a particular AI system for the worse.

See also: Business continuity; Change control; Model drift; Portability

Endpoint [C] — A network-accessible API path where requests are sent. Endpoint scope (public vs. private, authentication) is a key security factor.

See also: Authentication; Security

Environmental impact [R] — The energy consumption, carbon emissions, and resource use associated with AI training and inference. Environmental concerns increasingly appear in ESG reporting and procurement criteria; some regulations require disclosure.

See also: Compute; Sustainability; Training

Epoch — One complete pass through the entire training dataset during model training. Affects training cost and model behavior.

See also: Batch size; Hyperparameter; Training

Ethical AI — A broad term describing efforts to develop and use AI in ways aligned with stated values (e.g., fairness, transparency, accountability, privacy, safety). The term is used in governance frameworks and policy statements and is not a single technical standard.

See also: AI governance; Fairness; Responsible AI; Transparency

EU AI Act [R] — The European Union's primary AI regulation using risk-based categorization and obligations for providers and deployers. Key terms (provider, deployer, GPAI, high-risk, prohibited practices) affect compliance strategy for EU-facing deployments.

See also: Deployer; GPAI; High-risk AI system; Prohibited AI practices; Provider

Evaluation (evals) [C, T] — Repeatable testing measuring model/system quality, safety, robustness, and compliance with requirements. Evals support governance, regulatory compliance, and reasonable care; define who runs them and how results are handled.

See also: Benchmark; Monitoring; Red teaming; Safety evaluation

Evasion attack — An adversarial attack crafting inputs to cause misclassification of malicious content as benign. Evasion attacks can bypass content moderation and security filters; relevant to security representations.

See also: Adversarial attack; Content filtering; Robustness

Excessive agency [R] — A risk where AI systems are given more autonomy, permissions, or capabilities than necessary for their intended function. OWASP Top 10 for LLMs identifies excessive agency as a key vulnerability; least privilege principles apply.

See also: Agentic AI; Least privilege; Tool permissions

Execution — In the CORE framework, the dataflow connecting components from input to output, documenting how data travels through the system. Multiple execution paths may exist based on routing logic, conditional branches, or error handling. Documenting execution paths is needed for compliance with safety-by-design and privacy-by-design regulations as well as explainable AI.

See also: Audit; Components; CORE; Logging; Operations; Resources; XAI

Executive Orders on AI [R] — A series of U.S. presidential executive orders addressing artificial intelligence policy, with significant shifts between administrations. Key orders include: President Trump's 2019 order on "Maintaining American Leadership in Artificial Intelligence" (later codified in the National AI Initiative Act of 2020); President Biden's October 2023 Executive Order 14110 on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," which established reporting requirements for frontier models, mandated red-teaming, defined "dual-use foundation model," and directed creation of the AI Safety Institute; and President Trump's January 2025 orders revoking EO 14110 and replacing it with "Removing Barriers to American Leadership in Artificial Intelligence," which shifted policy toward deregulation and innovation. A December 2025 order sought to establish federal preemption of state AI laws. Despite the rescission of EO 14110, terminology it introduced (such as "dual-use foundation model" and "red team") remains in common use, and voluntary frameworks like the NIST AI RMF developed pursuant to it continue to be referenced in procurement and governance. Executive orders bind federal agencies but do not directly regulate private parties; however, they influence federal procurement requirements, agency enforcement priorities, and industry standards.

See also: Dual-use foundation model; NIST AI RMF (AI Risk Management Framework); Red teaming

Exfiltration — Unauthorized extraction of sensitive data from a system, including via model outputs or tool calls. Exfiltration risk is central to confidentiality and privacy; address via least privilege, prompt injection defenses, and logging.

See also: Data leakage; Prompt injection; Tool permissions

Explainability [C, R] — The ability to describe why a model produced a particular output in understandable terms. Explainability is often requested in regulated decisions. Due to limited understanding of exactly which factors lead to a particular output, explainability may be low. “Thinking” traces from reasoning models may be helpful for understanding why a model could generate a particular output, but they do not accurately represent what actually went into any particular decision.

See also: Explainable AI; Interpretability; Right to explanation

Explainable AI (XAI) — Techniques and methods designed to make AI model behavior more understandable to humans. XAI encompasses various approaches with different fidelity-complexity tradeoffs; claims often specify which methods are used and their limitations. Be careful not to confuse with X.ai, the AI provider associated with the microblogging platform X.

See also: Explainability; Interpretability

Export controls [R] — Laws restricting export of certain technologies (including AI hardware and software) to certain countries or parties. Export controls can affect model/compute sourcing, cross-border hosting, and M\&A diligence.

See also: Compute; Dual-use

Extraction attack — An adversarial attack seeking to extract training data, model weights, or system prompts through queries. Extraction attacks can expose confidential information and IP; relevant to security testing and rate limiting design.

See also: Data privacy attack; Model extraction; Prompt leakage

F

F1 score — A metric combining precision and recall, useful when classes are imbalanced. If a vendor promises "performance," clarify which metric matters (accuracy vs. F1 vs. recall) and on what dataset.

See also: Accuracy; Benchmark; Precision; Recall

Fairness [L, R] — The principle that AI systems should not produce unjustified differential outcomes for different groups. Fairness has multiple technical definitions (e.g. demographic parity, equalized odds, individual fairness) that can conflict with each other. A system cannot simultaneously satisfy all fairness definitions in most real-world scenarios. Contracts and governance documents should specify which definition applies and how it will be measured.

See also: Algorithmic discrimination; Bias; Disparate impact

Fair use [L] — A U.S. copyright doctrine permitting certain uses of copyrighted material without permission. Fair use is central to AI training data litigation; the doctrine's application to AI is actively being litigated.

See also: Copyright infringement (AI context); Training data; Transformative use

False negative — An instance incorrectly classified as negative when the true label is positive. False negative rates matter for safety-critical applications (missed fraud, missed medical conditions).

See also: False positive; Precision; Recall

False positive — An instance incorrectly classified as positive when the true label is negative. False positive rates affect user experience and can create liability (wrongful denials, false accusations).

See also: False negative; Precision; Recall

Feature — An individual measurable property of data used as input to a model (e.g., age, location, purchase history, medical codes). Feature choice affects model behavior; in regulated contexts, using protected characteristics (or close proxies) as features may trigger heightened review or restrictions depending on the use case and jurisdiction. In older models, features were frequently chosen by data scientists; in modern large models, features are usually discovered as part of the training process and may not be explicit.

See also: Bias; Feature engineering; Training data

Feature engineering — The process of selecting, transforming, and creating features for model training. Feature engineering choices can introduce bias or encode protected characteristics indirectly.

See also: Bias; Feature; Training

Federated learning — A training approach where model updates happen on distributed devices without centralizing raw data. Federated learning may be used as a privacy-enhancing technique.

See also: Data minimization; Privacy-enhancing technology; Training

Feedback loop [C] — A cycle where model outputs influence future training data or model behavior. Feedback loops can convert inference data into training data; contracts often address whether and how feedback is used.

See also: Monitoring; Service improvement

Few-shot prompting [T] — Providing a small number of examples in the prompt to guide model behavior without changing weights. Few-shot prompts act like operational policy; treat prompt libraries as sensitive assets with change control.

See also: In-context learning; Prompting; Zero-shot prompting

Field of use restriction [C] — A license term limiting where, how, or for what purposes a model may be used. Field-of-use restrictions are common in model licenses and can create compliance and enforcement issues.

See also: AUP; License grant; Open weights

Fine-tuned model [C] — A model produced by fine-tuning a base model on specific data. Determine ownership, permitted reuse, and confidentiality rules for fine-tuned artifacts and training datasets.

See also: Adapter; Fine-tuning; Model artifact

Fine-tuning [C, T] — Additional training of a pre-trained model on specialized data to customize behavior. Fine-tuning raises the sharpest IP and confidentiality issues: data rights, artifact ownership, and training set retention.

See also: Adapter; LoRA (Low-Rank Adaptation); Pre-training; Transfer learning

Floating-point operations (FLOPs) — A measure of computational work; used to quantify training and inference costs. FLOP thresholds appear in regulatory definitions (e.g. the EU AI Act and AI Executive Orders); relevant to compute cost negotiations.

See also: Compute; Training

Foundation model [C, R, T] — A large, general-purpose model trained on broad data that can be adapted to many tasks. Foundation model sourcing and licensing affect compliance posture; clarify whether vendor provides proprietary, open-weights, or wrapper.

See also: Base model; GPAI; LLM; Open weights

Frontier model [R] — An AI model at or near the cutting edge of capabilities, typically characterized by training compute, parameter count, and emergent abilities that may pose novel risks not present in less capable systems. Regulatory frameworks have attempted to formalize this concept using compute thresholds: President Biden's EO 14110 defined "dual-use foundation model" at 10^{26} FLOPs; the EU AI Act presumes "systemic risk" for GPAI models trained above 10^{25} FLOPs. Though EO 14110 was rescinded, compute thresholds remain relevant to export controls, international coordination, and voluntary industry commitments. The threshold for "frontier" shifts continuously—today's frontier becomes tomorrow's baseline.

See also: Dual-use foundation model; Executive Orders on AI; Export controls; Foundation model; GPAI; Systemic risk

Function calling [T]

See also: Tool calling (function calling)

Fundamental Rights Impact Assessment (FRIA) [R] — An assessment required under the EU AI Act for deployers of high-risk AI systems used to evaluate individuals, examining potential impacts on fundamental rights including privacy, non-discrimination, human dignity, and access to justice. FRIAs must be completed before first use and updated when circumstances materially change. Unlike DPIAs (which focus on data protection), FRIAs address broader rights impacts and require consideration of specific affected populations.

See also: Deployer; EU AI Act; High-risk AI system; Risk assessment

G

General Data Protection Regulation (GDPR) [R] — The EU's comprehensive data protection framework governing processing of personal data. In AI, GDPR concepts commonly implicated include lawful basis, transparency, data minimization, purpose limitation, security, cross-border transfers, and rights (e.g., access, deletion, objection); applicability depends on the processing context and the roles of controller/processor.

See also: Automated decision-making; Controller / processor; Personal data

Generalization — A model's ability to perform well on new, unseen data beyond its training set. Generalization relates to reliability; poor generalization can undermine product claims and cause failure on edge cases.

See also: Evaluation (evals); Overfitting; Robustness

General-purpose AI (GPAI) [R] — AI designed for broad applicability across many tasks and contexts; a category in the EU AI Act. GPAI terminology appears in EU AI Act with specific obligations for GPAI providers, especially those with systemic risk.

See also: EU AI Act; Foundation model; Systemic risk

Generated content — Content produced by a generative model (text, images, audio, video, code). "Generated content" is often used interchangeably with "Output," though contracts may define these differently.

See also: Generative AI (GenAI); Output

Generation [C] — Using a trained model to process new inputs and create new, unknown outputs. Unlike classification, recommendation, or extraction systems, generation is not constrained to predefined categories; the space of possible outputs is effectively unbounded. Generative responses will frequently include information provided as part of the prompt or from the training inputs.

See also: Generative AI (GenAI); Hallucination; Inference

Generative Adversarial Network (GAN) [T] — A generative model architecture using competing generator and discriminator networks. GANs powered early deepfake technology; largely superseded by diffusion models but still relevant for understanding synthetic media.

See also: Deepfake; Generative AI (GenAI); Synthetic data

Generative AI (GenAI) [C, L] — Models that generate new content (text, code, images, audio, video) rather than only classifying inputs. Generative outputs raise distinctive risks: hallucination, defamation, IP infringement, confidentiality leakage, deepfakes.

See also: Diffusion model; Hallucination; LLM

Governance, Risk, and Compliance (GRC) [C] — Term describing frameworks for managing governance policies, enterprise risk management, and regulatory compliance. Traditional GRC relies on organizational controls (such as written policies, training, procedures, and attestations) that work because humans read, understand, and follow them. In contrast, written policies do not constrain an AI model; the policy must be translated into technical controls such as guardrails, system prompts, tool permissions, and monitoring that govern actual system behavior. Effective AI governance requires mapping organizational controls (which govern humans who build and oversee AI) to technical controls (which govern what AI systems can do). Organizations with mature GRC functions can accelerate AI governance, but AI governance requires enforcement mechanisms beyond those designed for human compliance.

See also: AI governance; Controls; CORE; Guardrails; Risk assessment

GPAI

See also: General-purpose AI

Gradient descent — An optimization method that adjusts model weights in small steps toward lower prediction error. Think of it as rolling a ball downhill: the algorithm repeatedly moves weights in whatever direction reduces the loss function. Gradient descent determines both training speed and whether the model converges on useful patterns.

See also: Backpropagation; Loss function; Training

Graphics Processing Unit (GPU) [C, T] — Specialized hardware used to accelerate AI training and inference. GPU availability affects costs, vendor lock-in, and export control considerations.

See also: Compute; Export controls

Groundedness — A RAG quality metric measuring whether outputs are supported by the provided context rather than fabricated. Groundedness is key to RAG reliability claims; ungrounded outputs may involve fabrication.

See also: Grounding; Hallucination

Grounding [C, T] — Constraining outputs to specified sources (retrieved documents, databases) rather than model patterns alone. Grounding improves defensibility and reduces hallucination risk; requires controls over source corpus and retrieval logs.

See also: Citation; Hallucination

Ground truth — The best available correct labels or answers used to train or evaluate a system. Ground truth can be disputed in subjective domains; define labeling standards and dispute processes.

See also: Annotation; Evaluation (evals); Labeling

Group Relative Policy Optimization (GRPO) — A training technique for reasoning models that optimizes based on relative performance within groups of responses. Introduced in late 2024 and used in many frontier models.

See also: Direct Preference Optimization (training); Reasoning model

Guardrails [C] — Technical and procedural controls limiting unsafe or undesired behavior. Guardrails are often the practical basis for safety representations and compliance strategies.

See also: Content filtering; HITL; Safety policy

H

Hallucination [C, L, T] — Plausible-sounding output that is wrong, unsupported, or fabricated, including fake citations. Hallucinations drive malpractice, consumer protection, and litigation risk; mitigation combines grounding, evals, and human review.

See also: Grounding; Verification

High-risk AI system [R] — An AI system category with heightened regulatory obligations due to impact on safety or fundamental rights. High-risk categorization drives documentation, testing, human oversight, and post-market monitoring duties under the EU AI Act.

See also: Conformity assessment; EU AI Act; Monitoring

HITL

See also: Human-in-the-loop

Hosting [C] — The environment where a model or AI system runs (cloud, on-prem, managed service), including compute, networking, and storage. Hosting affects security controls, data residency, latency, and incident response responsibilities.

See also: Cloud provider; Data residency; On-prem deployment; Security

Human evaluation [T] — Assessment of AI outputs by human reviewers for quality, safety, or compliance. Document human evaluation methods to support quality claims and manage reviewer exposure to sensitive content.

See also: Evaluation (evals); Labeling; Red teaming

Human-in-the-loop (HITL) [C, R] — A control where a human reviews, approves, or edits outputs before use. HITL can be a decisive risk mitigator and regulatory requirement for high-stakes decisions.

See also: Approval workflow; Automated decision-making; Autonomy level

Hybrid search — Retrieval that combines traditional keyword search with vector/semantic search to improve recall and precision. Hybrid search is common in enterprise RAG implementations.

See also: Semantic search; Vector database

Hyperparameter [T] — A configuration setting chosen by engineers (not learned), such as learning rate or temperature. Hyperparameters affect performance and reproducibility; relevant in disputes about model version changes.

See also: Parameter; Temperature; Training

I

Image generation — AI creation of images from text prompts or other inputs. Image generation raises copyright questions regarding training data and outputs, deepfake and NCII concerns, and trademark issues when generated images depict recognizable brands or logos.

See also: Diffusion model; Generative AI (GenAI); Text-to-image

Incident response [C] — Processes for detecting, responding to, and recovering from AI system failures or security events. AI incident response plans often address model-specific scenarios such as hallucination-caused harm, prompt injection breaches, and unexpected behavior changes; contracts commonly specify notification and cooperation obligations.

See also: Business continuity; Monitoring; Security

In-context learning — A model's ability to adapt behavior within a prompt using provided examples, without weight changes. Because in-context learning allows prompts and retrieved documents to materially change behavior, prompt content is often treated as part of the controlled system.

See also: Context window; Few-shot prompting; Prompting

Indemnity [C] — A contractual promise by one party to defend and/or reimburse the other for specified third-party claims (often including costs). In AI contexts, indemnities commonly address IP claims, data protection incidents, and misuse claims, with scope turning on definitions of “Input,” “Output,” and “Training” and on compliance with use restrictions.

See also: IP indemnity; Limitation of liability; Warranty

Indirect prompt injection [T] — A prompt-injection technique where malicious instructions are embedded in data the system retrieves or processes (e.g., web pages, emails, tickets, documents) rather than in the user’s direct prompt. This is a common risk in RAG and tool-enabled agent systems.

See also: Connector; Prompt injection; Tool calling (function calling)

Inference [C, T] — Using a trained model to process new inputs and create outputs. This is the operational phase, as distinguished from training. Inference is distinguished from training by the way in which data is handled: training uses input data to adjust model weights, leading to possible issues with memorization, while inference only uses input data transiently to produce outputs. Inference is also distinguished from generation; a classification system performs inference, but the scope of possible outputs is limited. In contrast, generative systems may reproduce input or training data. Contracts often define "inference" and "training" separately and impose different restrictions on each.

See also: Generation; Latency; Training

Inference-time scaling [T] — Computational resources used during inference, particularly the extended processing in reasoning models that "think" before responding. Inference-time compute affects cost and latency; reasoning models may use substantially more compute per query than standard models.

See also: Compute; Inference; Reasoning model

Input Data [C] — Content provided to an AI system as input, including prompts and context documents. Input Data often contains sensitive information, so contracts commonly define whether it can be retained, logged, or used for service improvement.

See also: Confidential information; Logging; User prompt

Insecure output handling — A security risk where AI outputs are used in downstream systems without adequate validation or escaping (e.g., injecting generated content into code, HTML, SQL, or commands). This can convert hallucinations or malicious outputs into system actions.

See also: Prompt injection; Security; Tool calling (function calling)

Instruction hierarchy (prompt precedence) [T] — The priority order that determines which instructions a model follows when multiple sources exist (commonly: system prompt \> developer instructions \> user prompt \> tool outputs/ retrieved text). Understanding precedence is central to prompt injection and guardrail design.

See also: Developer prompt; Prompt injection; System prompt; User prompt

Instruction tuning — Fine-tuning a model to follow instructions and engage in dialogue. Instruction-tuned models (often called "chat" models) behave differently than base models, which is relevant when assessing capabilities and limitations.

See also: Chat model; Fine-tuning

Intellectual property (IP) [C, L] — Legal rights in creations of the mind, including patents, copyrights, trademarks, and trade secrets. AI raises novel IP questions across all categories, with training data, outputs, and model architecture each presenting distinct issues requiring separate analysis.

See also: Copyright; IP indemnity; Patent; Trade secret

Intended purpose [R] — Under the EU AI Act, the use for which an AI system is intended by the provider as specified in required documentation, including instructions for use, promotional materials, and technical specifications.

See also: EU AI Act; High-risk AI system; Off-label use; Provider; Technical documentation

Intended use [C, R] — The purpose, context, and conditions under which a model or AI system is designed and evaluated to operate (e.g., internal drafting vs automated decisions; healthcare vs general productivity). “Intended use” is commonly used to scope warranties, safety controls, and regulatory obligations.

See also: AUP; Evaluation (evals); Off-label use; Risk assessment

Interpretability — The degree to which a human can understand how a model produces its outputs. Interpretability is stronger than explainability and implies genuine understanding of internal mechanisms; truly interpretable models are often less capable than black-box alternatives.

See also: Black box; Explainability; XAI

IP

See also: Intellectual property

IP indemnity [C] — A contractual promise to defend and compensate for intellectual property infringement claims. AI IP indemnities vary widely in scope (training data vs. output), exceptions (user modifications, combinations), and conditions (cooperation, control of defense), requiring careful negotiation.

See also: Copyright; Indemnity; Training data

ISO/IEC 27001 [C] — An international standard for information security management systems (ISMS). References to ISO 27001 often appear in security questionnaires and vendor contracts as evidence of a structured security program.

See also: Security; Security addendum; SOC 2

ISO/IEC 42001 [C, R] — An international standard for AI management systems (AIMS), ISO 42001 follows the Plan-Do-Check-Act structure of other ISO management system standards (like ISO 27001 for information security), making it integrable with existing compliance programs. The standard covers AI system lifecycle management, risk assessment and impact evaluation, data governance, third-party supplier oversight, and 38 specific controls in its annexes. Organizations already certified to ISO 27001 can leverage significant structural overlap.

See also: AI governance; Audit; ISO/IEC 27001; NIST AI RMF (AI Risk Management Framework); Risk assessment; SOC 2

J

Jailbreak [C] — An attempt to bypass a model's safety restrictions through crafted prompts. Jailbreak resistance is part of safety evaluation, and successful jailbreaks may trigger AUP violations, incident response obligations, and potential liability for resulting harms.

See also: Adversarial attack; Prompt injection; Safety evaluation

Jamba — A family/name used for architectures combining elements of transformers with state space models (SSMs) to improve efficiency for long sequences. The term is used in technical discussions about model architecture and inference cost.

See also: Context window; Transformer

JSON mode — A setting constraining model outputs to valid JSON format. Structured output modes like JSON mode improve reliability and parsing for automated workflows, reducing integration errors and enabling programmatic processing of AI outputs.

See also: Function calling; Structured output

K

Key-Value cache (KV) — Memory storing prior context during inference to enable efficient generation. KV cache size affects context window limits and inference cost, which is relevant when understanding capacity constraints and pricing models.

See also: Context window; Memory; Transformer

Knowledge base [C] — A curated collection of documents or data used for retrieval in RAG systems. Knowledge base contents determine output quality and risk exposure, making access control, accuracy verification, and update processes important governance considerations.

See also: Connector; Grounding

Knowledge distillation [T]

See also: Distillation

Knowledge transfer — The process of moving know-how, documentation, and operational understanding from one team or vendor to another (e.g., during vendor transitions, M\&A, or outsourcing). In AI deployments, knowledge transfer may include model documentation, evals, runbooks, and data pipelines.

See also: Business continuity; Documentation; Portability

L

Labeling

See also: Annotation

Language model [T] — A model that assigns probabilities to sequences of tokens; modern large language models (LLMs) are language models scaled and trained for broad capabilities. “Language model” can also refer to smaller or domain-specific models.

See also: Large Language Model; Token; Transformer

Large Language Model (LLM) [C, T] — A neural network trained on extensive text data to understand and generate language. "LLM" is often used interchangeably with "AI" in business contexts, though technically it refers to text-focused models; modern LLMs are increasingly multimodal.

See also: Foundation model; Generative AI (GenAI); Small Language Model; Transformer

Large Reasoning Model (LRM) — An LLM specifically trained for complex reasoning, often using extended chain-of-thought processing. Reasoning models like o1 and DeepSeek-R1 represent a capability shift and may be more reliable for complex analysis, though they have different cost and latency profiles.

See also: Chain of Thought (CoT); Inference-time scaling; Reasoning model

Latency [C, T] — The time delay between sending a request and receiving a response. Latency affects user experience and may be specified in SLAs; high latency can undermine real-time use cases and is often tested under realistic load conditions.

See also: Inference; Service Level Agreement / Service Level Objective; Throughput

Latent space — The abstract multi-dimensional space encoding learned representations of data. Embeddings exist in latent space, and understanding this concept helps explain how AI systems represent and compare semantic meaning.

See also: Embedding; Representation learning

Lawful basis [R] — Under GDPR, one of six legal grounds that must exist before processing personal data: consent, contractual necessity, legal obligation, vital interests, public task, or legitimate interests. AI systems often rely on legitimate interests (requiring a balancing test) or consent (requiring clear, specific, freely-given agreement). Training on personal data, inference processing, and service improvement uses may each require separate lawful basis analysis.

See also: Consent; Personal data; Privacy; Purpose limitation

Least privilege [C] — The security principle of granting only the minimum permissions necessary for an actor (user, service, or agent) to perform its task. In agentic and tool-enabled systems, least-privilege permissioning and scoped tool access are common controls to reduce the impact of errors, abuse, or prompt injection.

See also: Access control; Security; Tool permissions

Lethal Trifecta — A security vulnerability pattern identified by Simon Willison occurring when an AI agent simultaneously possesses three capabilities: (1) access to private or sensitive data, (2) exposure to untrusted content, and (3) the ability to communicate externally. When all three capabilities are present, prompt injection attacks can cause the agent to access private data and transmit it to an attacker. The Lethal Trifecta has been demonstrated against major products including Microsoft 365 Copilot, ChatGPT, Google Gemini, Slack, and GitHub Copilot. Because prompt injection remains an unsolved problem, the primary defense is to ensure AI systems never combine all three capabilities simultaneously.

See also: Agentic AI; Agents Rule of Two; Exfiltration; Prompt injection; Tool permissions

License compatibility — Whether multiple licenses (e.g., open source licenses, model licenses, and proprietary licenses) can be complied with simultaneously when components are combined or distributed. Incompatibilities can arise from obligations such as copyleft, attribution, field-of-use limits, or downstream restrictions.

See also: Copyleft license; Open source

License compliance [C] — Adhering to terms of software and model licenses. AI systems often combine multiple licensed components with different terms, requiring tracking and satisfying all applicable obligations to avoid infringement claims.

See also: Copyleft license; Open source

License grant [C] — The specific permissions conveyed by a license. Model license grants vary widely in scope, field of use restrictions, sublicensing rights, and modification permissions, requiring careful examination for each intended use.

See also: Field of use restriction; IP; Open weights

Limitation of liability [C] — Contract terms that limit damages (e.g., caps, exclusions of consequential damages, and carve-outs). In AI service agreements, parties often allocate risk differently across categories such as IP claims, security incidents, confidentiality, and misuse; the negotiated structure varies by use case and regulatory exposure.

See also: Contract; Indemnity; Warranty

Litigation hold [L] — A process to preserve relevant information when litigation or an investigation is reasonably anticipated. In AI systems, holds may apply to logs, prompts, outputs, tool call records, and model/version artifacts.

See also: E-discovery; Logging; Retention

LLM [T]

See also: Large Language Model

Logging [C] — Recording system activity including prompts, outputs, and operational data. Logging is essential for audit and debugging but creates privacy and confidentiality exposure; contracts commonly define what is logged, who can access logs, and retention periods.

See also: Monitoring; Records retention; Usage data / telemetry

LoRA (Low-Rank Adaptation) [C, T] — A parameter-efficient fine-tuning technique that trains small adapter weights rather than full model weights. LoRA adapters are common deliverables in custom AI projects, so ownership, portability to other base models, and confidentiality terms are often stated explicitly.

See also: Adapter; Fine-tuning; Parameter-efficient fine-tuning

Loss function — A mathematical formula that measures how wrong the model's predictions are. The choice of loss function determines what the model optimizes for; a model trained to minimize one type of error may perform poorly by other measures. This is relevant when evaluating whether a model was designed appropriately for its intended use: a model optimizing for average accuracy may systematically fail on minority cases.

See also: Gradient descent; Training

LRM

See also: Large Reasoning Model

M

Machine learning (ML) — A subset of AI where systems learn patterns from data rather than following explicit rules. ML is the technical foundation of modern AI, and understanding that models learn statistical patterns (rather than "knowing" facts) helps assess capabilities and limitations.

See also: Deep learning; Neural network; Training

Machine Unlearning [R] — Techniques intended to reduce or remove the influence of specific data from a trained model without full retraining. Approaches and effectiveness vary; some providers offer deletion workflows or re-training processes, but the scope of guarantees, documentation, and verification methods differs across implementations.

See also: Algorithmic Disgorgement (Model Deletion); Guardrails; Right to be Forgotten

Mamba [T] — A selective state space model architecture offering efficient processing of long sequences. Mamba represents an alternative to transformer architecture for some applications, particularly those requiring very long context windows.

See also: Jamba; Transformer

Massive Multitask Language Understanding (MMLU) — A benchmark testing model knowledge across many academic subjects. MMLU scores are commonly cited in model marketing, but benchmarks don't guarantee real-world performance and may not reflect capabilities on domain-specific tasks.

See also: Benchmark; Evaluation (evals)

MCP

See also: Model Context Protocol

Membership inference [T] — An attack determining whether specific data was used in training. Membership inference can reveal training data composition, which is relevant to privacy claims, trade secret disputes, and copyright litigation discovery.

See also: Data privacy attack; Memorization; Model inversion

Memorization [C, L] — A model's tendency to reproduce training data verbatim rather than generalizing. Memorization creates copyright and privacy exposure and is the mechanism behind extraction attacks; the degree of memorization varies by model and data frequency. Usually the result of overfitting to repeatedly-seen training data.

See also: Copyright; Data leakage; Extraction attack; Overfitting

Memory — Stored state used across interactions (e.g., conversation history, user preferences, task state). Memory can be ephemeral (within a context window) or persistent (stored and retrieved later), and it can raise retention, privacy, and confidentiality considerations.

See also: Context window; Logging; Personal data; Retention

Metadata — Data describing other data, including creation dates, sources, and processing history. AI-generated content may lack authentic metadata or have synthetic metadata, which is relevant to evidence authentication and content provenance disputes.

See also: Content provenance; Data provenance

Mixture of Experts (MoE) [T] — A neural network architecture where only a subset of the model's parameters, called "experts", are activated for each input. MoE allows models to have very large total parameter counts while keeping inference costs manageable: a model with 400 billion parameters might activate only 50 billion for any given query.

See also: Architecture; Compute; Inference; Parameter

MLOps — Practices for deploying and maintaining ML systems in production, including monitoring, versioning, and updates. MLOps maturity affects reliability, reproducibility, and change control; assess vendor MLOps practices as part of due diligence.

See also: DevOps; Model registry; Monitoring

Model [C] — The trained artifact (weights and architecture) that processes inputs to produce outputs. In contracting, “model” is commonly distinguished from “AI system” because model licensing and ownership are often negotiated separately from system-level concerns like hosting, data handling, and integration.

See also: AI system; Model artifact; Weights

Model artifact [C] — Tangible outputs of model development and deployment, such as weights, checkpoints, adapters, fine-tuned models, training logs, evaluation results, and prompt templates. These artifacts are often treated as valuable intellectual property and can be addressed in development, licensing, and confidentiality terms.

See also: Adapter; Checkpoint; Weights

Model capability — The range and level of tasks a model can perform reliably, often evidenced through evaluations, benchmarks, and red teaming results. Capability statements can affect intended-use scoping and governance classification.

See also: Benchmark; Capabilities; Evaluation (evals); Intended use

Model card — Structured documentation describing a model's intended use, performance characteristics, and limitations. Model cards support due diligence and can undermine marketing claims if disclosed limitations conflict with vendor representations.

See also: Documentation; Evaluation (evals); Technical documentation

Model collapse [C] — The degenerative process where a model trained on synthetic (AI-generated) data eventually loses quality, diversity, and connection to reality. As the internet fills with AI-generated content, "organic" human-generated training data becomes a premium asset; contracts may need to specify the ratio of synthetic versus organic data to support data quality warranties.

See also: Data provenance; Synthetic data; Training data

Model compression — Techniques reducing model size while maintaining performance, including quantization and distillation. Compressed models may behave differently than originals, and compression is a form of modification that may require license analysis.

See also: Distillation; Edge deployment; Quantization

Model Context Protocol (MCP) — An open source standard protocol for connecting AI models to external data sources and tools. MCP addresses the execution layer: how an agent calls tools and retrieves data. MCP enables interoperability between different AI systems and data sources, with implications for data access control, logging, and vendor lock-in. The use of MCP (usually described in terms of an “MCP server”) almost always implies that an AI system will be given access to some Resource.

See also: Connector; CORE; Resources; Tool calling (function calling)

Model deprecation

See also: End-of-life / deprecation

Model drift — Changes in model behavior over time due to updates, data changes, or environmental shifts. Drift drives change control clauses in contracts and can create unexpected compliance failures in validated workflows.

See also: Change control; Monitoring; Version pinning

Model escrow / artifact escrow [C] — An arrangement where model weights, training data, adapters, system prompts, or other AI artifacts are deposited with a third-party escrow agent for release to the customer upon specified trigger events.

See also: Adapter; Business continuity; Trade secret; Vendor lock-in; Weights

Model extraction — Attempting to recreate a model or approximate its behavior through repeated API queries. Model extraction is often prohibited in AUPs and relevant to trade secret protection; rate limiting and query monitoring serve as partial defenses.

See also: Distillation; Extraction attack; Rate limiting; Trade secret

Model improvements [C] — Enhancements to models developed during or after initial deployment, including new weights, adapters, and prompts. Ownership of improvements is a recurring dispute, particularly when customer data, feedback, or funding contributed to the work.

See also: Adapter; Fine-tuned model; License grant

Model inversion — Attacks inferring training data characteristics from model outputs. Model inversion is relevant to privacy assurances and security controls when sensitive data may have influenced training, even indirectly.

See also: Data privacy attack; Membership inference; Memorization

Model license [C] — The terms governing how a model (weights, API access, or hosted service) may be used, including restrictions, attribution, field-of-use limits, and downstream obligations. Model licenses differ from open source software licenses and can be bespoke.

See also: Field of use restriction; License grant; Open source; Open weights

Model registry — A system tracking model versions, metadata, approvals, and deployment status. Registries support governance and auditability by providing a single source of truth for what models exist, where they're deployed, and who approved them.

See also: Change control; MLOps; Model drift

Model Risk Management (MRM) [C, R] — A governance discipline for validating, monitoring, and controlling model use, particularly common in financial services under SR 11-7 guidance. MRM programs often impose documentation, independent validation, and change control obligations that apply to AI systems.

See also: AI governance; Evaluation (evals); Monitoring

Model supply chain — The set of upstream components and processes used to build and operate a model or AI system (datasets, code, weights, third-party models, tools, connectors, hosting, and subprocessors). Supply chain analysis is used for security, IP provenance, and compliance.

See also: Security; Subprocessor; Training data

Model update [C] — A change to model weights, tuning, safety settings, prompts, or configuration. Updates can introduce drift or unexpected behavior changes; contracts may require notice, version pinning options, and regression testing for critical workflows.

See also: Change control; Model drift; Version pinning

Moderation [C] — Processes (automated and human) detecting and managing disallowed content or behavior. Moderation intersects with platform liability, AUP enforcement, and employment law; it also creates recordkeeping requirements and content reviewer welfare considerations.

See also: AUP; Content filtering; Safety policy

Monitoring [C, R] — Ongoing observation of a model or AI system in production (e.g., quality, safety events, latency, errors, drift, abuse signals). Monitoring outputs are used in incident response, governance reporting, and contract performance management.

See also: Data drift; Incident response; Model drift; SLA/SLO

Multi-agent system (MAS) — An architecture where multiple AI agents collaborate or compete to accomplish tasks. Multi-agent systems create complex liability and attribution challenges because harm may result from emergent interactions rather than any single agent's action.

See also: Agentic AI; Orchestration

Multimodal [T]

See also: Multimodal model

Multimodal model [C, T] — A model accepting and/or producing multiple content types (text, images, audio, video). Multimodal capabilities expand privacy and IP risk through processing of faces, voices, and biometrics, and may trigger additional regulatory obligations.

See also: Biometric data; Computer vision; LLM

Multi-tenancy [C] — Cloud architecture where multiple customers share infrastructure with logical separation. Multi-tenancy raises confidentiality concerns for AI systems; confirm isolation controls for prompts, embeddings, fine-tuned models, and logs.

See also: Data segregation; SaaS (Software as a Service); Security

Multi-tenant — A service architecture in which multiple customers share underlying infrastructure and software while remaining logically separated. Multi-tenancy affects data segregation, logging, and security controls.

See also: Access control; Data segregation; SaaS (Software as a Service)

N

Narrow AI — AI designed for specific tasks rather than general intelligence. All current AI systems are narrow AI, regardless of marketing claims; this is relevant when assessing vendor capability representations.

See also: Foundation model

Natural Language Processing (NLP) — AI techniques for understanding, analyzing, and generating human language. NLP encompasses many AI applications from classification to generation, with legal issues varying significantly by task type.

See also: Language model; LLM; Text classification

NCII

See also: Non-Consensual Intimate Images

Neural network [T] — A computational architecture that processes information through layers of interconnected nodes, with each connection carrying a learned weight. Neural networks are not deterministic rule-based systems. Rather, they learn statistical patterns from training data, which explains both their capabilities and their tendency to produce confident-sounding errors. The term "neural" is a historical metaphor; these systems do not function like biological brains.

See also: Deep learning; Transformer; Weights

NIST AI RMF (AI Risk Management Framework) [R] — A voluntary framework published by the National Institute of Standards and Technology providing guidance for managing AI risks throughout the system lifecycle. The framework is organized around four core functions: Govern (establishing accountability, policies, and culture), Map (understanding context, stakeholders, and potential impacts), Measure (assessing and tracking risks through evaluation and monitoring), and Manage (prioritizing and responding to identified risks). The framework is widely referenced in U.S. federal procurement, various AI Executive Orders, sector guidance, and enterprise customer requirements. NIST also published the Generative AI Profile (AI RMF 600-1) addressing risks specific to generative AI systems. Alignment with the NIST AI RMF is frequently requested in vendor assessments and can support reasonable care arguments, though "alignment" is self-assessed and does not involve certification or audit.

See also: AI governance; Executive Orders on AI; ISO/IEC 42001; Risk assessment; Trustworthy AI

NLP

See also: Natural Language Processing

Non-Consensual Intimate Images (NCII) [L, R] — Intimate images shared or generated without consent. AI-generated NCII is addressed by a growing set of criminal, civil, and platform policy regimes; obligations and remedies vary by jurisdiction and may depend on intent, distribution, and notice-and-takedown processes.

See also: Content filtering; Deepfake; Prohibited AI practices

Non-deterministic — A system property where the same input may produce different outputs across runs. Most generative AI systems are non-deterministic by default due to sampling strategies, floating-point computation variations, and infrastructure differences. Non-determinism affects auditability, testing reproducibility, and user expectations; controls like temperature settings can enforce more deterministic behavior when needed.

See also: Deterministic; Reproducibility; Sampling; Temperature

Notice [C, R] — A communication provided to another party or to individuals (e.g., privacy notices, product disclosures, contractual notices of changes or incidents). "Notice" requirements are often defined by contract or applicable law, and may include timing and content requirements.

See also: Disclosure; Incident response; Transparency

No training on our data [C] — A procurement requirement that a provider not use customer content (e.g., prompts, files, outputs, or connected data) to train, fine-tune, or otherwise improve its models beyond providing the contracted service. Implementations vary and may distinguish between model training, human review, debugging, safety monitoring, and logging/retention practices.

See also: Service improvement; Training; Usage data / telemetry

O

Observability [C] — The ability to understand system behavior through logs, metrics, and traces. Observability supports audit, debugging, and incident response; contracts commonly define what telemetry is available and whether customers can access it.

See also: Incident response; Logging; Monitoring

Off-label use — Using an AI system for purposes beyond its intended or permitted use. Off-label use can affect contractual rights (including warranties and indemnities), compliance posture, and safety assumptions because the system may not have been evaluated or controlled for the new context.

See also: AUP; Field of use restriction; Intended use

On-prem deployment [C] — Running AI systems on customer-controlled infrastructure rather than vendor cloud. On-prem deployment affects data residency compliance, security responsibility allocation, update mechanisms, and support arrangements.

See also: Data residency; Edge deployment; SaaS (Software as a Service)

Open source [C, R] — A term most precisely applied to software code distributed under an open source license. In AI discussions, "open source" is sometimes used more loosely to refer to publicly available weights, datasets, or systems with permissive access, which may not match established open source definitions.

See also: Model license; Open source software; Open weights

Open source software (OSS) [C] — Software distributed under licenses permitting use, modification, and redistribution with varying conditions. AI systems often incorporate open source components; compliance requires tracking all licenses and satisfying their respective obligations.

See also: Copyleft license; License compliance

Open weights [C] — Models where trained weights are publicly released, but not necessarily under open source licenses. "Open weights" is distinct from "open source"; many open weight licenses include field of use restrictions and may not meet the Open Source Initiative definition.

See also: Field of use restriction; Foundation model; License grant

Operations — In the CORE framework, the actions that components perform on external Resources or on data flowing through an AI system. Operations include data transformations (summarization, classification, generation), resource interactions (reading from databases, calling external APIs, writing outputs), and control functions (filtering, routing, logging).

See also: Components; CORE; Execution; Resources; Tool calling (function calling)

Optimization — Methods to improve a model or system's performance, cost, latency, or resource use (e.g., quantization, caching, batching, prompt compression). Optimization choices can affect accuracy, safety behavior, and reproducibility.

See also: Key-Value cache; Latency; Quantization

Orchestration — Coordinating multiple AI components, tools, and workflows to accomplish complex tasks. Orchestration layers make consequential decisions about tool selection and sequencing, requiring oversight, logging, and clear accountability.

See also: Agentic AI; Multi-agent system; Tool calling (function calling)

Output [C] — Content produced by an AI system in response to inputs. Outputs may contain confidential information from inputs or retrieval, create IP ownership questions, or cause harm through inaccuracy; contracts commonly define ownership, permitted uses, and retention.

See also: Generated content; Hallucination; IP

Overfitting — When a model performs well on training data but poorly on new data because it memorized specific examples rather than learning generalizable patterns. Overfitting explains why demo performance may not match production results, and frequently leads to memorization and subsequent regeneration of training material.

See also: Evaluation (evals); Generalization; Memorization; Training

OWASP Top 10 for LLMs [R] — A security framework identifying top vulnerabilities in LLM applications, including prompt injection, insecure output handling, and excessive agency. The OWASP Top 10 provides a standard reference for AI security requirements and risk assessment.

See also: Excessive agency; Prompt injection; Security

P

Parameter [T] — A learned value in a neural network that influences outputs; parameter count (often in billions) indicates model size. Parameter count is often cited as a capability proxy, but actual performance depends on architecture, training data, and post-training, not just size.

See also: Hyperparameter; Model; Weights

Parameter-efficient fine-tuning (PEFT) [T] — Fine-tuning techniques that modify only a subset of parameters, reducing computational cost. PEFT techniques like LoRA produce smaller artifacts than full fine-tuning, but ownership and portability of these artifacts still requires contractual clarity.

See also: Adapter; Fine-tuning; LoRA (Low-Rank Adaptation)

Patent [L] — A form of intellectual property that can protect inventions meeting statutory requirements (e.g., novelty, non-obviousness, utility), subject to jurisdiction-specific eligibility rules. In AI contexts, patents may cover model architectures, training techniques, and system implementations.

See also: Intellectual property; Prior art; Trade secret

PEFT

See also: Parameter-efficient fine-tuning

Performance — How well a model or system meets task objectives and operational requirements (accuracy, latency, robustness, cost, safety). In contracting, performance is often expressed as SLAs/SLOs, acceptance criteria, and evaluation benchmarks tied to intended use.

See also: Benchmark; Evaluation (evals); Reliability; SLA/SLO

Perplexity [T] — A metric measuring how well a language model predicts text, with lower values indicating better performance. Perplexity is a technical quality metric primarily useful for model comparison rather than legal analysis.

See also: Benchmark; Evaluation (evals)

Personal data [R] — Information relating to an identified or identifiable natural person. AI systems processing personal data trigger privacy obligations under GDPR, state laws, and sector regulations; prompts, outputs, and training data may all contain personal data.

See also: Data subject; Privacy

Personally Identifiable Information (PII) — Information that can identify an individual, with the precise definition varying by jurisdiction. PII in training data, prompts, or outputs creates compliance obligations; detection and filtering controls are important safeguards.

See also: De-identification; Personal data; Privacy

PET

See also: Privacy-enhancing technology

Plugin — A software extension enabling additional functionality in AI systems, such as web browsing, code execution, or database access. Plugins expand AI system capabilities and risks by accessing external resources, and may require separate permissions and security review.

See also: Connector; Tool calling (function calling)

Portability [C] — The ability to move an AI workflow, data, or model artifacts from one environment or vendor to another (e.g., switching model providers, migrating vector databases). Portability can turn on data formats, licensing terms, and access to artifacts.

See also: Business continuity; Model artifact; Open weights; Vendor lock-in

Post-market monitoring [R] — An ongoing obligation under the EU AI Act requiring providers of high-risk AI systems to collect and analyze data on system performance and compliance after deployment. Post-market monitoring is distinct from general operational monitoring; it requires a documented plan proportionate to the system's risks, must feed into the provider's quality management system, and triggers reporting and corrective action obligations when issues are detected.

See also: High-risk AI system; Monitoring; Provider; Quality Management System; Serious incident

Post-training — Modifications to a model after pre-training to change behavior or improve usefulness and safety (e.g., instruction tuning, preference optimization, safety tuning, or distillation). Post-training often changes model characteristics and can affect evaluation results, safety properties, and documentation baselines.

See also: Fine-tuning; Pre-training

Precision — A metric measuring, of items classified as positive, the fraction that were truly positive. High precision means fewer false positives, which is important for applications where false accusations or unnecessary interventions are costly.

See also: F1 score; False positive; Recall

Pre-training — Initial training of a model on large-scale data before any customization or alignment. Pre-training data provenance is the primary driver of copyright and privacy exposure for foundation models.

See also: Foundation model; Post-training; Training

Prior art [L] — Publicly available information that can be relevant to assessing patentability (or invalidity) of an invention. In AI, prior art may include papers, open source code, model cards, and public model releases.

See also: Open source software; Patent; Publication

Privacy [R] — Protection of personal information from unauthorized collection, use, or disclosure. AI creates novel privacy challenges through inference capabilities, behavioral profiling, and potential memorization of training data; multiple legal frameworks apply.

See also: Personal data; Privacy-enhancing technology

Privacy by design [R] — Embedding privacy protections into systems from the start rather than as an afterthought. Privacy by design is a GDPR principle requiring consideration of privacy throughout AI development, not just at deployment.

See also: Data minimization; Privacy

Privacy-enhancing technology (PET) [R] — Technical means of mitigating privacy risks, including encryption, differential privacy, federated learning, and secure enclaves. PETs may support privacy claims but have limitations; understand what protections they actually provide in practice.

See also: Differential privacy; Federated learning; Privacy

Probabilistic — A fundamental characteristic of how generative AI models operate: outputs are generated by sampling from learned probability distributions over possible responses rather than by executing logical rules or retrieving stored facts. Even when configured for deterministic operation, a probabilistic model is selecting the statistically most likely output based on training patterns, not computing a provably correct answer. This distinction explains why models can be confidently wrong (hallucination), why explanations of "reasoning" may be post-hoc rationalizations, and why traditional software warranties and performance guarantees require adaptation for AI systems.

See also: Deterministic; Explainability; Hallucination; Neural network; Sampling

Prohibited AI practices [R] — AI applications banned under regulatory frameworks such as the EU AI Act. The EU AI Act prohibits certain uses including social scoring, real-time remote biometric identification in public spaces, and emotion recognition in workplaces and schools, with severe penalties for violations.

See also: Biometric data; EU AI Act; High-risk AI system

Prompt [C, T] — The input text or instructions provided to an AI system to generate a response. Prompts often contain sensitive business information or personal data; contracts commonly define whether they are logged, retained, or used for improvement.

See also: Input Data; System prompt; User prompt

Prompt engineering [T] — The practice of crafting prompts to achieve desired AI behaviors and outputs. Prompt engineering creates valuable know-how, and prompt libraries may constitute confidential information or trade secrets worth protecting.

See also: Few-shot prompting; Prompting; System prompt

Prompting [T] — The practice of supplying instructions, context, and examples to a model to elicit desired outputs. Prompting includes system/developer prompts and user prompts and is often an alternative to fine-tuning for adapting model behavior.

See also: Few-shot prompting; Prompt engineering; System prompt

Prompt injection [C, R, T] — An attack in which malicious input causes a model or agent to ignore intended instructions or perform unintended actions (e.g., by overriding system/developer prompts or by exploiting tool integrations). Prompt injection can occur directly via user input or indirectly via retrieved content and is treated as a security risk in many threat models.

See also: Adversarial attack; Jailbreak; Security

Prompt leakage [T] — Unintended disclosure of system prompts or instructions through model outputs. System prompts may contain confidential business logic, competitive information, or security controls; leakage can expose sensitive operational details.

See also: Confidential information; Extraction attack; System prompt

Prompt library [C, T] — A collection of tested prompts for specific tasks or use cases. Prompt libraries represent operational controls and potential trade secrets that is often treated as governed assets with change control and access restrictions.

See also: Change control; Confidential information; Prompt engineering

Provider [R] — A party that develops or places an AI system on the market, as defined in the EU AI Act. Provider vs. deployer role allocation determines documentation, conformity assessment, and incident reporting duties under the regulation.

See also: AI system; Deployer; EU AI Act

Pseudonymization [R] — Replacing direct identifiers with tokens while retaining the ability to re-link data under safeguards. Pseudonymized data is generally still personal data under GDPR and similar regimes; do not treat it as equivalent to anonymization.

See also: De-identification; Personal data; Privacy

Publication — Making information publicly available (e.g., papers, open source repositories, model releases, technical reports). Publications can be relevant to prior art, marketing claims, and transparency commitments.

See also: Model card; Open source software; Prior art

Purpose limitation [R] — A privacy principle requiring data be used only for specified, explicit, and legitimate purposes. Purpose limitation applies when vendors want to reuse prompts, outputs, or logs for improvement; permitted purposes is often defined explicitly.

See also: Data minimization; DPA; Service improvement

Q

Quality Assurance (QA) — Processes ensuring AI systems meet quality standards before and during deployment. QA processes support reliability claims and reasonable care arguments; document testing methodologies and acceptance criteria.

See also: Acceptance criteria; Evaluation (evals); Testing

Quality Management System (QMS) [R] — A documented system of policies, procedures, and processes required of providers under the EU AI Act to ensure consistent compliance with regulatory requirements throughout AI system development and operation. QMS obligations for high-risk AI systems include risk management procedures, data governance, technical documentation practices, post-market monitoring, vulnerability identification, incident reporting protocols, and recordkeeping.

See also: AI governance; High-risk AI system; ISO/IEC 42001; Post-market monitoring; Provider; Vulnerability / CVE

Quantization [C, T] — Reducing numerical precision of model weights to decrease memory requirements and speed inference. Quantization can change model behavior in subtle ways; treat quantized models as different versions requiring separate validation for regulated deployments.

See also: Edge deployment; Model compression; Performance

R

Rate limiting [C] — Controls capping API usage to manage load, prevent abuse, and control costs. Rate limits affect service usability, are often disclosed in SLAs, and also serve as security controls against model extraction and denial-of-service attacks.

See also: Abuse; Service Level Agreement / Service Level Objective

ReAct (Reason \+ Act) — An agent prompting pattern combining reasoning traces with action execution in an interleaved manner. ReAct is a common pattern in agentic AI systems that makes agent decision-making somewhat more transparent.

See also: Agentic AI; Chain of Thought (CoT); Tool calling (function calling)

Reasoning model [C] — An AI model specifically trained for complex reasoning, often with extended "thinking" time before responding. Reasoning models like o1 and DeepSeek-R1 represent a capability advancement for analysis tasks but have different cost, latency, and transparency profiles.

See also: Chain of Thought (CoT); Inference-time scaling; Large Reasoning Model

Recall — A metric measuring, of truly positive items, the fraction that were correctly identified. High recall is critical for safety applications where missing positives is costly, such as fraud detection or medical screening.

See also: F1 score; False negative; Precision

Records of Processing Activities (ROPA) [R] — Documentation required under GDPR Article 30 cataloging an organization's personal data processing activities, including purposes, data categories, recipients, transfers, retention periods, and security measures. AI systems should be reflected in ROPA entries, with particular attention to training data processing, inference logging, and any cross-border transfers to model providers.

See also: Controller / processor; Cross-border data transfer; Data retention; Personal data

Records retention [C] — Policies governing how long business records are kept before deletion. AI logs may become business records subject to retention requirements, litigation holds, and deletion rights; align retention periods with legal obligations.

See also: Data retention; E-discovery; Logging

Recovery Point Objective / Recovery Time Objective (RPO / RTO) [C] — RPO specifies the maximum acceptable data loss measured in time (e.g., "no more than 4 hours of data"); RTO specifies the maximum acceptable downtime before service restoration. These metrics are standard in disaster recovery planning but require special attention for AI systems. Contracts should specify RPO/RTO for each critical component (models, vector stores, knowledge bases, configuration) and clarify whether RTO includes re-indexing and validation time, not just data restoration.

See also: Availability; Business continuity; Service Level Agreement / Service Level Objective; Vector database

Red teaming [C, R] — Adversarial testing by dedicated teams attempting to find system vulnerabilities and failure modes. Red teaming is referenced in the AI Executive Order and EU AI Act; it supports safety claims and regulatory compliance when properly documented.

See also: Adversarial attack; Evaluation (evals); Safety evaluation

Reinforcement Learning (RL) — A training approach where the model learns by receiving rewards or penalties for its outputs rather than by studying labeled examples. A primary benefit of RL is that in some circumstances it allows synthetic data or self-play to be used in the place of human-labeled data. RL is the foundation for RLHF and is used in game-playing AI and robotics; understanding RL helps explain how models learn to follow instructions.

See also: Reward function; Training

Reinforcement Learning from AI Feedback (RLAIF) — A variant of RLHF using AI-generated feedback rather than human feedback to reduce cost and scale. RLAIF may have different alignment properties than human feedback and is often combined with Constitutional AI approaches.

See also: Alignment; Constitutional AI

Reinforcement Learning from Human Feedback (RLHF) [C, T] — Training models using human preference data to improve alignment with human values and instructions. RLHF is the dominant technique for aligning LLMs and is key to understanding how models are made to follow instructions and refuse harmful requests.

See also: Alignment; Post-training; Reward model

Reinforcement Learning with Verifiable Rewards (RLVR) — Training using automatically verifiable outcomes such as correct code execution or valid mathematical proofs. RLVR is used for training reasoning models on tasks with objectively checkable answers.

See also: Reasoning model; Reinforcement Learning

Reliability [C, R] — Consistency and dependability of system behavior over time, including correctness under expected conditions and stability across updates. Reliability is influenced by drift, stochasticity, evaluation coverage, and operational monitoring.

See also: Evaluation (evals); Model drift; Monitoring; Temperature

Representation learning — Learning useful internal representations of data that can transfer to multiple downstream tasks. Foundation models learn representations that encode semantic meaning, enabling capabilities like semantic search and few-shot learning.

See also: Embedding; Foundation model; Transfer learning

Reproducibility [C] — The ability to obtain consistent results from the same inputs across runs. Reproducibility is important for testing, audit, and debugging; non-determinism can be controlled through temperature and seed settings but may affect output quality.

See also: Evaluation (evals); Non-deterministic; Temperature

Re-ranking — A retrieval step (often in RAG) that reorders candidate results using a second model (e.g., a cross-encoder or an LLM) to improve relevance. Re-ranking can affect what content is presented to the generation model and therefore affects grounding and auditability.

See also: Answer relevance; Retrieval

Resources — In the CORE framework, external assets that an AI system accesses but does not control, such as data, third-party APIs, file systems, knowledge bases, and external services. Resources exist outside the system boundary but are invoked during execution. Resource mapping is essential for data sovereignty compliance, confidentiality protection, and contractual obligation tracking.

See also: Components; Connector; CORE; Data residency; Operations; Tool calling (function calling)

Responsible AI [R] — Principles and practices for developing and deploying AI ethically, safely, and in accordance with human values. Responsible AI frameworks inform governance programs, procurement criteria, and regulatory expectations.

See also: AI governance; AI safety; Ethical AI

Retention — Policies and technical practices governing how long data and logs are kept and when they are deleted. In AI systems, retention can apply to prompts, outputs, embeddings, retrieval indexes, and tool call records.

See also: Data governance; Deletion; Litigation hold; Logging; Zero Data Retention

Retrieval [T] — The process of selecting and returning relevant information from a corpus or database (often using keyword or semantic search) to support tasks such as RAG. Retrieval quality affects grounding, hallucination rates, and completeness.

See also: Re-ranking; Semantic search; Vector database

Retrieval-Augmented Generation (RAG) [C, T] — An architecture where the system retrieves relevant documents and uses them as context for generation, grounding outputs in specific sources. RAG improves accuracy and reduces hallucination but introduces document access logging, confidentiality, and injection risks that require governance.

See also: Grounding; Knowledge base; Vector database

Reverse engineering [C] — Analyzing a system to understand its design, extract components, or replicate functionality. Reverse engineering restrictions commonly appear in AI terms of service; they may conflict with legitimate security research and interoperability needs.

See also: AUP; Distillation; Model extraction; Trade secret

Reward function — In reinforcement learning, the function that assigns a numeric score (“reward”) to behaviors, guiding the model toward preferred outcomes. Reward function design influences aligned behavior and can encode tradeoffs.

See also: Alignment; Reinforcement Learning

Reward model — A model trained to predict human preferences, used to guide RL training by scoring candidate outputs. Reward model quality directly affects alignment effectiveness and the behaviors the final model learns to exhibit.

See also: Alignment

Right of Publicity [L, R] — A set of state-law (and sometimes statutory) rights controlling commercial use of a person’s name, image, likeness, or voice. In AI contexts, right-of-publicity issues can arise with voice cloning, deepfakes, and digital replicas.

See also: Consent; Deepfake; Digital Replica

Right to be Forgotten [R] — A term commonly referring to deletion rights under certain privacy frameworks (notably in EU law), including circumstances where individuals can seek removal of personal data from search results or other systems. How it applies to AI training data and model artifacts depends on the legal framework and technical implementation.

See also: Deletion; Machine Unlearning

Right to explanation [R] — A data subject's right to understand the logic of automated decisions affecting them. GDPR Article 22 and various U.S. state laws provide explanation rights for significant automated decisions, though the required depth of explanation remains debated.

See also: Automated decision-making; Explainability

Risk assessment [C, R] — Systematic evaluation of AI system risks, potential harms, and likelihood of occurrence. Risk assessment is required by the EU AI Act for high-risk systems and expected by governance frameworks; it is often documented and periodically updated.

See also: AI governance; High-risk AI system; NIST AI RMF (AI Risk Management Framework)

Robustness [C] — A system's ability to maintain performance under varying conditions, distribution shifts, and adversarial inputs. Robustness claims often specify what conditions and attack types were tested; general robustness guarantees are difficult to provide.

See also: Adversarial attack; Evaluation (evals); Reliability

S

SaaS (Software as a Service) [C] — Cloud-hosted software accessed over the internet on a subscription basis. Most AI products are delivered as SaaS, with key issues including data handling practices, logging, subprocessor use, and service level commitments. Warning: a SaaS may not be a "product" for purposes of product liability.

See also: Multi-tenancy; Service Level Agreement / Service Level Objective

Safe harbor [C, R] — A legal provision shielding parties from liability or enforcement when they meet specified conditions. In the AI context, safe harbors appear in several forms: some state AI laws (such as Colorado's AI Act) provide safe harbors for organizations following recognized frameworks like the NIST AI RMF; Section 230 provides platform immunity that may apply to certain AI-generated content (although the scope is contested); and contractual safe harbors may limit liability when parties follow agreed procedures. Safe harbor protection typically requires documented, good-faith implementation rather than mere assertion of alignment.

See also: AI governance; ISO/IEC 42001; Limitation of liability; NIST AI RMF (AI Risk Management Framework)

Safety evaluation [C, R, T] — Testing specifically focused on identifying unsafe, harmful, or disallowed behaviors. Safety evaluations support governance, regulatory compliance, and reasonable care arguments; they may be required by regulators for certain AI applications.

See also: Evaluation (evals); Guardrails; Red teaming

Safety policy [C] — Rules defining prohibited behaviors for an AI system, implemented through training, prompts, and filtering. Safety policy affects AUP enforcement, duty-of-care arguments, and user expectations; it is often documented and traceable to implemented controls.

See also: AUP; Guardrails; Moderation

Sampling — Selecting outputs probabilistically from the model's predicted distribution rather than always choosing the highest-probability option. Sampling contributes to output variability; deterministic modes may be needed for audit and reproducibility requirements.

See also: Non-deterministic; Temperature; Top-p (nucleus) sampling

Sandboxing [C] — Running code or processes in a constrained environment that limits access to sensitive resources. Sandboxing is a key control for agentic systems executing code, supporting security representations and limiting blast radius of failures.

See also: Least privilege; Security; Tool calling (function calling)

Scalability [C] — The ability to maintain performance and availability as usage grows. Scalability affects SLA reliability and business continuity; capacity constraints during high demand can create significant operational and reputational issues.

See also: Performance; Rate limiting; Service Level Agreement / Service Level Objective

Scaling law — Empirical relationships showing that model performance improves predictably with increased size, data, and compute. Scaling laws drive investment in larger models and help explain capability improvements, though they don't guarantee specific abilities.

See also: Compute; Foundation model; Parameter

Secrets exposure — Unauthorized disclosure of secrets (API keys, credentials, tokens, confidential prompts) through prompts, logs, tool outputs, or model behavior. This is discussed in AI security, incident response, and vendor controls.

See also: Access control; Logging; Prompt injection; Security

Secure development lifecycle (SDLC) [C] — Processes integrating security practices throughout software development. SDLC commitments may appear in security addenda and support reasonable-security arguments in the event of incidents.

See also: Development practices; Security

Security [C] — Protection of systems and data from unauthorized access, use, disclosure, or destruction. AI systems require security controls appropriate to the sensitivity of data processed and decisions made; common frameworks like SOC 2 provide baseline standards.

See also: Access control; Encryption; SOC 2

Security addendum [C] — A contract attachment specifying security requirements, controls, and audit rights. AI security addenda often address model-specific concerns including prompt injection defenses, data isolation, logging access, and incident notification.

See also: Contract; DPA; Security

Self-attention [T] — The attention mechanism applied within a single sequence to capture relationships between different positions. Self-attention is the core mechanism enabling transformers to understand context and relationships in language.

See also: Attention mechanism; Transformer

Semantic search [T] — Search based on meaning and intent rather than exact keyword matching, typically using embeddings to find semantically similar content. Semantic search powers RAG retrieval; understanding its behavior helps assess retrieval quality and potential failure modes.

See also: Embedding; Vector database

Sensitive Personal Information (SPI) [R] — Categories of personal data treated as especially sensitive under some privacy regimes (e.g., precise geolocation, health data, biometric identifiers, financial data, or data about minors). In AI systems, SPI handling often drives additional controls for collection, processing, retention, access, and disclosures.

See also: Biometric data; Personal data; Privacy

Serious incident [R] — Under the EU AI Act, an incident or malfunction of a high-risk AI system that directly or indirectly causes death, serious damage to health or property, or serious and irreversible disruption of critical infrastructure. Serious incidents trigger mandatory reporting to market surveillance authorities. Incident classification and reporting procedures should be integrated into the provider's QMS.

See also: EU AI Act; High-risk AI system; Incident response; Post-market monitoring; Provider; Quality Management System

Service credits [C] — Contractual remedy providing credits against future fees when a vendor fails to meet SLA commitments, typically calculated as a percentage of monthly fees based on the severity and duration of the failure. Service credits are the standard remedy in SaaS agreements.

See also: Availability; Contract; Limitation of liability; Service Level Agreement / Service Level Objective

Service improvement [C] — Using customer data, including prompts, outputs, and feedback, to improve AI services. Service improvement rights are contentious in negotiations; clearly define whether and how customer content can be used, and provide opt-out mechanisms.

See also: No training on our data; Training; Usage data / telemetry

Service Level Agreement / Service Level Objective (SLA/SLO) [C] — Contractual commitments on service performance metrics such as availability, latency, and throughput. AI SLAs often address model-specific issues including response time variability, rate limits, and model deprecation notice.

See also: Availability; Latency; Uptime

Shadow AI — Unauthorized use of AI tools by employees outside official channels and governance processes. Shadow AI increases confidentiality breach risk, privilege waiver concerns, and compliance exposure; it requires both policy controls and technical measures to address.

See also: Acceptable use; AI governance

Similarity search [T] — Finding items semantically close to a query by comparing embeddings in vector space. Similarity thresholds affect retrieval quality and is often documented for regulated deployments where retrieval completeness matters.

See also: Cosine similarity; Vector database

SLA/SLO

See also: Service Level Agreement / Service Level Objective

Slop — Pejorative term for low-quality, mass-produced AI-generated content, typically created with minimal human oversight or input.

See also: AUP; Content provenance; Generated content; Generative AI (GenAI); Synthetic media; Watermarking

Small Language Model (SLM) [C, T] — A language model with fewer parameters (typically \<10B) designed for efficiency, lower latency, and often local/edge deployment. Unlike larger models, SLMs can often run on consumer hardware (laptops, phones) without sending data to a cloud provider, changing the privacy and security risk profile.

See also: Distillation; Edge deployment; Quantization

SOC 2 [C] — An audit framework and report on controls for security, availability, processing integrity, confidentiality, and privacy. SOC 2 reports are standard due diligence for AI vendors; confirm the report scope includes AI-relevant systems and controls.

See also: Audit; Security

Software Bill of Materials (SBOM) [C] — An inventory of software components, dependencies, and their versions used in a system. SBOMs support open source license compliance, vulnerability management, and supply chain security; they are increasingly requested in enterprise procurement.

See also: License compliance; Open source; Supply chain security

Source attribution — Linking generated outputs to the documents or data used to produce them. Attribution supports defensibility and user trust but can be incorrect or fabricated; validate citation mechanisms through testing.

See also: Citation; Grounding

SPI

See also: Sensitive Personal Information

Statement of Work (SOW) [C] — A contract section or attachment describing specific services, deliverables, and acceptance criteria. AI SOWs often specify model boundaries, evaluation metrics, change control procedures, and artifact ownership.

See also: Acceptance criteria; Contract; Deliverables

State Space Model (SSM) [T] — A neural architecture modeling sequences using continuous state representations, offering efficient processing of long sequences. SSMs like Mamba are alternatives to transformers for applications requiring very long context windows or efficient inference.

See also: Context window; Mamba; Transformer

Structured output [C] — Constraining AI outputs to a defined schema such as JSON, XML, or specific formats for reliable parsing. Structured output improves auditability, reduces parsing errors, and enables automated processing in enterprise workflows.

See also: Function calling; JSON mode

Subprocessor [C] — A third party engaged by a vendor to process data on its behalf. Subprocessor lists, notification of changes, and flow-down of data protection obligations are central provisions in DPAs governing AI services.

See also: Cloud provider; Controller / processor; DPA

Substantial modification [R] — Under the EU AI Act, a change to a high-risk AI system after initial placing on the market that affects compliance with regulatory requirements or alters the system's intended purpose. Changes to training data, model architecture, or safety controls may constitute substantial modifications even if commercial framing remains unchanged.

See also: Change control; Deployer; EU AI Act; High-risk AI system; Model update; Provider

Substantial similarity [L] — The legal standard for copyright infringement based on whether works are sufficiently similar in protected expression. Substantial similarity analysis is central to AI copyright litigation.

See also: Copyright; Fair use

Summarization — Condensing longer content into shorter form while preserving key information. AI summarization may miss important details, introduce errors, or change emphasis; users often verify accuracy for legal and business-critical matters.

See also: Hallucination; Truncation

Supervised learning — Machine learning from labeled examples that map inputs to correct outputs. Supervised learning requires labeled training data, raising data rights, annotation labor, and quality control considerations.

See also: Labeling; Machine learning; Training

Supply chain security [C] — Protecting against risks from third-party components, data sources, and service providers. AI supply chains include foundation models, training datasets, open source libraries and other components, as well as third-party cloud infrastructure; each introduces potential vulnerabilities.

See also: Data poisoning; Open source

Sustainability — Environmental and resource considerations of AI development and deployment (energy use, water use, hardware lifecycle). Sustainability may be discussed in procurement, ESG reporting, and policy debates about compute-intensive systems.

See also: Compute; Datacenter

Synthetic data — Data generated by a computer process (possibly an AI system) rather than collected from real-world sources. Synthetic data can reduce privacy concerns and augment training sets but may not accurately represent real-world distributions or edge cases.

See also: Data augmentation; Model collapse; Privacy; Training data

Synthetic media — AI-generated content including images, audio, video, and text created to resemble authentic content. Synthetic media raises authenticity, deepfake, evidence authentication, and misinformation concerns; provenance and detection tools are evolving. A number of laws require the disclosure of synthetic media.

See also: Content provenance; Deepfake; Generative AI (GenAI)

System card — Documentation describing a complete AI system's design, intended use, capabilities, limitations, and safety measures. System cards provide more complete context than model cards by covering the full deployed system, not just the underlying model.

See also: Architecture; CORE; Documentation; Model card; Technical documentation

Systemic risk [R] — Risk that AI system failures or misuse could have wide-ranging negative impacts on society, the economy, or critical infrastructure. The EU AI Act imposes additional transparency and evaluation obligations on GPAI models posing systemic risk.

See also: EU AI Act; GPAI; High-risk AI system

System prompt [C, T] — Instructions provided to the model that frame its role, capabilities, and constraints, typically hidden from end users. System prompts contain business logic and operational controls that may be confidential; prompt leakage is a security concern.

See also: Confidential information; Prompt; Prompt leakage

T

Technical documentation [R] — Formal documentation of AI system design, development, testing, and deployment required by regulatory frameworks. The EU AI Act requires comprehensive technical documentation for high-risk systems; retain documentation for compliance and liability defense.

See also: Conformity assessment; Model card; System card

Temperature [C, T] — A parameter controlling output randomness, where lower values produce more deterministic and focused outputs. Temperature settings affect reproducibility, creativity, and consistency; document settings for regulated or audited workflows.

See also: Hyperparameter; Non-deterministic; Sampling

Tensor Processing Unit (TPU) [T] — Google's custom-designed AI accelerator hardware optimized for neural network operations. TPUs are alternatives to GPUs with different availability, pricing, and vendor dependencies.

See also: Compute

Terms of Service [C] — Contract terms governing use of a product or service (often online), including license grants, restrictions, disclaimers, limitation of liability, and incorporated policies such as an AUP. Terms of Service are commonly relevant to authorization, enforcement, and dispute resolution.

See also: Acceptable Use Policy; Contract; Limitation of liability; Warranty

Testing — Methods for assessing whether a model or system meets requirements and behaves safely and reliably (e.g., unit tests, red teaming, evals, regression tests). Testing results are often used in acceptance, governance, and incident response.

See also: Acceptance criteria; Change control; Evaluation (evals); Red teaming

Text classification — A common inference task where a model assigns one or more labels to text (e.g., spam/not spam, topic tagging, sentiment). Classification models may be built with traditional ML, deep learning, or LLM prompting.

See also: Classification; Inference; NLP; Supervised learning

Text-to-image — AI generation of images from text descriptions or prompts. Text-to-image systems raise copyright questions about training data and outputs, trademark concerns, and deepfake risks.

See also: Diffusion model; Generative AI (GenAI); Image generation

Threat model — An analysis identifying potential attackers, attack vectors, assets at risk, and security boundaries for a system. Threat models inform security requirements, testing priorities, and control design; they should be documented and updated as systems evolve.

See also: Adversarial attack; OWASP Top 10 for LLMs; Prompt injection; Risk assessment; Security

Throughput [C, T] — The rate of processing, typically measured in requests per second or tokens per second. Throughput affects system capacity and cost; it may be specified in SLAs and is often tested under realistic load conditions.

See also: Latency; Rate limiting; Service Level Agreement / Service Level Objective

Token [C, T] — The unit of text that models actually process. Tokens are not words: common words may be single tokens, but less common words are split into pieces ("contract" might be one token; "indemnification" might be three). A rough approximation is 0.75 words per token for English. Token counts determine API costs, which are typically priced per token, and they define context window limits. Understanding tokenization helps explain why non-English text and technical terminology often perform worse–they require more tokens to represent the same meaning.

See also: Context window; Rate limiting; Tokenization

Tokenization [T] — The process of converting text into tokens that the model can process. Tokenization affects how different languages, technical content, and special characters are handled, potentially causing issues with non-English text or domain-specific terminology. In contrast to vectorization, tokenization is a reversible, direct translation of the input.

See also: Context window; Token; Vectorization

Tool calling (function calling) [C, T] — A mechanism where the model outputs structured instructions to invoke external functions, APIs, or services. Tool calling is a key inflection point for risk because it enables AI systems to take real-world actions; permissions, logging, and approval workflows are essential controls.

See also: Agentic AI; Least privilege

Tool output injection — A risk where untrusted tool outputs (e.g., web content, search results, emails) introduce malicious instructions or data that affect subsequent model behavior. This is closely related to indirect prompt injection in tool-enabled systems.

See also: Indirect prompt injection; Prompt injection; Tool calling (function calling)

Tool permissions [C] — Controls specifying what actions and resources AI agents can access. Permissions are primary controls for limiting potential harm from agentic systems; implement least privilege principles and require explicit authorization for sensitive operations.

See also: Agentic AI; Excessive agency; Least privilege

Top-k sampling [T] — A sampling method limiting next-token selection to the k highest-probability options. Top-k is one of several sampling parameters affecting output determinism; document settings for reproducibility requirements.

See also: Sampling; Temperature; Top-p (nucleus) sampling

Top-p (nucleus) sampling [T] — A sampling method limiting selection to tokens within a cumulative probability threshold. Like temperature, top-p affects output variability and is often documented for audit and reproducibility purposes.

See also: Sampling; Temperature; Top-k sampling

Trade secret [C] — Information deriving economic value from not being generally known and subject to reasonable secrecy efforts. Model weights, training data compositions, system prompts, and prompt libraries may qualify as trade secrets; logging and vendor access can undermine secrecy claims.

See also: Confidential information; IP; Weights

Training [C] — The process of adjusting model weights using data to improve performance on target tasks. Training definitions are central to negotiations about customer data use, distinguishing training from inference, evaluation, and service improvement.

See also: Fine-tuning; Post-training; Pre-training

Training data [C, L] — Data used to train or fine-tune a model, including text, images, code, and other content. Training data provenance drives copyright and privacy exposure for AI systems; it is central to indemnity negotiations and ongoing litigation.

See also: Copyright; Dataset documentation; IP indemnity

Transfer learning — Adapting a pre-trained model to new tasks rather than training from scratch, leveraging learned representations. Transfer learning underlies most commercial AI applications and drives questions about base model rights versus adaptation rights.

See also: Fine-tuning; Foundation model; Pre-training

Transformative use [L] — In copyright fair use analysis, whether a use adds new meaning, message, or purpose to the original work. Transformative use is a key factor in AI training data litigation; courts are actively deciding how the doctrine applies to machine learning.

See also: Copyright; Fair use; Training data

Transformer [C, T] — The neural network architecture underlying modern LLMs and most other frontier AI systems. Transformers process input by converting it to tokens, then using attention mechanisms to determine which parts of the input are relevant to each other. This architecture enables models to handle long-range dependencies in text, such as understanding that a pronoun in one sentence refers to a noun several paragraphs earlier. "Transformer-based" in vendor materials signals a model with LLM-like capabilities.

See also: Architecture; Attention mechanism; LLM

Transparency [R, L] — The degree to which information about an AI system is disclosed and understandable (e.g., intended use, data sources at a high level, limitations, evaluation results, safety controls). Transparency is commonly discussed in policy, procurement, and consumer protection contexts.

See also: Disclosure; Documentation; Model card; System card

Tree-of-thoughts prompting [T] — A prompting framework that explores multiple reasoning paths in a tree structure rather than a single chain. Tree-of-thoughts is an advanced technique for complex reasoning tasks requiring evaluation of alternatives.

See also: Chain of Thought (CoT); Prompting; Reasoning model

Truncation — Dropping content when inputs exceed context window limits or other constraints, often without explicit notification. Truncation can undermine reliability because users may not know the system ignored portions of their input; it creates risk for legal document review.

See also: Context window; Summarization; Token

Trusted execution environment — A hardware- or enclave-based environment designed to protect code and data during execution (e.g., isolating workloads from a host OS). TEEs are discussed in confidential computing and can be used to reduce exposure when processing sensitive data.

See also: Confidential computing; Encryption; Security

Trustworthy AI [R] — AI systems exhibiting characteristics including validity, reliability, safety, security, transparency, fairness, and privacy protection. The NIST AI RMF defines trustworthiness characteristics that inform governance programs and regulatory expectations.

See also: AI governance; NIST AI RMF (AI Risk Management Framework); Responsible AI

U

Unbounded consumption [R] — A vulnerability where AI systems consume excessive computational or financial resources without limits. OWASP Top 10 for LLMs identifies this as a key risk; rate limiting, budget controls, and resource monitoring are mitigations.

See also: OWASP Top 10 for LLMs; Rate limiting; Security

Unsupervised learning — Learning patterns from unlabeled data without explicit correct answers. Unsupervised techniques like clustering can create sensitive inferences about individuals or groups even without labeled protected attributes.

See also: Clustering; Embedding; Machine learning

Uptime [C] — The percentage of time a service is operational and accessible over a measurement period. Uptime commitments in AI SLAs often address model-specific failure modes; pair with incident response and communication terms.

See also: Availability; Service Level Agreement / Service Level Objective

Usage data / telemetry [C] — Operational data about service use, which may include prompts, outputs, timestamps, and performance metrics. Telemetry is often where service improvement occurs; contracts commonly define permitted collection, retention periods, and use restrictions.

See also: Data retention; Logging; Service improvement

User prompt [T] — The end-user's input to an AI system, as distinguished from system prompts set by developers. User prompts frequently contain personal or confidential information; define logging practices and provide appropriate privacy notices. In many cases, the user prompt is the proximate cause of the AI system output.

See also: Input Data; Logging; Privacy

V

Validation — Testing that a system meets its specified requirements and performs acceptably for intended uses. Validation is key to acceptance criteria and regulatory compliance; document validation methodology and results.

See also: Acceptance criteria; Evaluation (evals); Testing

Value alignment — Ensuring AI systems pursue goals and exhibit behaviors consistent with human values and intentions. Value alignment is a central AI safety concept; misalignment between system objectives and human values can cause harmful behavior even without adversarial attack.

See also: AI safety; Alignment

Variational Autoencoder (VAE) [T] — A generative model architecture that encodes data as probability distributions in latent space rather than fixed points. VAEs are used in some image generation systems and for learning compressed representations.

See also: Autoencoder; Generative AI (GenAI); Latent space

Vector [C, T] — A list of numbers representing an item in a mathematical space. Embeddings are not a direct translation of the item content; instead, vectors are a one-way transformation that encodes semantic meaning. Nevertheless, vectors derived from sensitive content may themselves be sensitive; assess whether embeddings constitute personal data or confidential information.

See also: Embedding; Latent space

Vector database [C, T] — A database optimized for storing embeddings and performing fast similarity search at scale. Vector databases often store representations of sensitive enterprise content; access control, encryption, and retention policies are critical security considerations.

See also: Embedding; Semantic search

Vectorization [T] — Converting inputs into numeric vector representations for model processing by representing their statistical values in a virtual high-dimensional space. Vectorization raises questions about whether derived representations retain the legal significance of source content, including personal data characteristics.

See also: Embedding; Tokenization; Vector

Vendor lock-in [C] — Dependence on a specific vendor that makes switching costly or difficult. AI lock-in can arise from proprietary formats, fine-tuned models, prompt libraries, and integrated workflows; consider portability of all assets when selecting vendors.

See also: Business continuity; Portability; Version pinning

Verification — Confirming that outputs, claims, or system behaviors are accurate and meet requirements. Human verification is a key control for managing hallucination risk in high-stakes applications; define verification requirements and responsibilities.

See also: Hallucination; Human-in-the-loop

Version pinning [C] — Locking to a specific model version to prevent unplanned behavior changes from updates. Pinning is important for validated and regulated workflows; contracts commonly address version availability, deprecation notice, and migration support.

See also: Change control; Model drift; SLA/SLO

Vision-language model [T] — A multimodal model that processes both visual inputs (images/video) and text, enabling tasks such as captioning, visual question answering, and document understanding.

See also: Computer vision; Multimodal model

Vulnerability / CVE [C] — A weakness in software, systems, or models that could be exploited to cause harm; CVE (Common Vulnerabilities and Exposures) is the standardized system for identifying and cataloging such weaknesses. Traditional software vulnerabilities in AI systems or components follow established CVE disclosure and patching processes–and these processes may be required for some systems due to regulations like the EU AI Act and Cyber Resilience Act (CRA). AI-specific vulnerabilities like prompt injection and jailbreaks are less standardized; MITRE ATLAS and OWASP Top 10 for LLMs provide emerging taxonomies but lack CVE-style universal identifiers. Contracts should address vulnerability disclosure timelines, patching obligations, and notification requirements for both software and model-level vulnerabilities. When evaluating vendor security posture, assess both traditional vulnerability management (CVE monitoring, patch cadence) and AI-specific security practices.

See also: Cyber Resilience Act; Incident response; OWASP Top 10 for LLMs; Prompt injection; Security

W

Warranty [C] — A contractual assurance about a product or service (e.g., performance, conformity to documentation, security controls, non-infringement). In AI agreements, warranties are frequently scoped by intended use, evaluation criteria, and exclusions for customer inputs, misuse, or model updates.

See also: Indemnity; Limitation of liability; Service Level Agreement / Service Level Objective

Watermarking [R] — Embedding detectable signals in AI-generated content to indicate its origin or enable provenance tracking. Watermarks can support authenticity verification but may be fragile or removable; do not overstate their reliability as detection mechanisms.

See also: Content provenance; Metadata

Weights [C] — The learned numerical parameters of a neural network that determine its behavior; the core model artifact. Weights are frequently treated as valuable trade secrets and key IP; access restrictions, licensing terms, and security controls are important diligence topics.

See also: Model; Open weights; Parameter

X

XAI

See also: Explainable AI

Z

Zero Data Retention (ZDR) [C] — A vendor commitment not to retain customer prompts, outputs, or associated data beyond the duration necessary to process the request and return a response. ZDR is often offered as an API option or enterprise tier feature to address confidentiality, privacy, and "no training on our data" concerns. However, ZDR policies vary significantly in scope: clarify whether ZDR covers abuse monitoring logs, trust and safety reviews, debugging data, error logs, metadata, and cached embeddings. Some vendors retain data briefly (e.g., 30 days) for abuse detection even under "zero retention" labels; others exclude certain content flagged for safety review. ZDR does not address what happened to data *before* the policy was enabled, nor does it prevent data exposure during transmission or processing.

See also: Confidential information; Data retention; Logging; No training on our data; Service improvement; Usage data / telemetry

Zero-shot prompting [T] — Prompting a model to perform a task without providing examples, relying entirely on the model's pre-existing capabilities. Zero-shot performance varies significantly by task; testing often verifies the model can reliably perform intended uses without examples.

See also: Few-shot prompting; In-context learning; Prompting