AI Technology Stack Glossary

The USCC AI Technology Stack Glossary is a plain-English reference guide designed to help Congress and the broader public better understand the complex technical concepts shaping today's technology and national security landscape. The glossary explains key terms related to semiconductors, semiconductor manufacturing, and artificial intelligence (AI), providing readers with a practical resource for navigating policy discussions, industry developments, and emerging technologies. The terms in this product have been sorted into relevant categories for ease of understanding, but please note that some terms may fall into more than one category. This product will be updated on a semi-regular basis with additional key terms and concepts.

Last updated on June 25, 2026

Hardware

Term	Definition
3-D Chip Stacking	A chipmaking process in which multiple chips are vertically stacked on top of each other, rather than laid out flat side-by-side. This design allows data to move more quickly between chips, improving performance and energy efficiency. It is especially useful for artificial intelligence (AI), high-performance computing, and other applications that require large amounts of data to be processed quickly.
Application-Specific Integrated Circuit (ASIC)	A computer chip designed to perform one specific task rather than a wide range of tasks. ASICs are built for a particular application, which usually makes them faster and more power-efficient for that job than a general-purpose chip such as a central processing unit (CPU).
Chip Packaging	The process of enclosing a finished semiconductor chip in a protective structure that allows it to connect to a circuit board, communicate with other components, and manage power, heat, and performance. Advanced chip packaging can also significantly improve computing performance and energy efficiency by enabling faster data transfer, denser integration of multiple chips, and reduced power loss between components.
Chiplet	A small chip built to do one part of a bigger computing job, such as processing, memory, or input/output. Several chiplets can be linked together inside one package so they work like a single larger chip.
Deep Ultraviolet (DUV) Lithography	A semiconductor manufacturing technology that uses short wavelengths of light (typically 193–248 nanometers [nm]) to print tiny circuit patterns onto silicon wafers (defined below). DUV lithography is used to produce a wide range of semiconductors, from mature-node chips to advanced logic and memory chips, but it is generally less capable than the newer extreme ultraviolet (EUV) lithography used for leading-edge semiconductor manufacturing. Because DUV and EUV systems rely on fundamentally different optical designs, light sources, and materials, existing DUV machines cannot simply be upgraded or converted into EUV machines.

Extreme Ultraviolet (EUV) Lithography	A semiconductor manufacturing technology that uses extremely short wavelengths of light to print tiny circuit patterns on silicon wafers at nanometer scale. Because EUV light can be absorbed by any form of matter (including air), the manufacturing process must operate in a near-perfect vacuum and be repeated dozens of times with extreme precision. Dutch firm ASML is the world's sole producer of EUV lithography machines, which are required to manufacture the most advanced chips that power today's cutting-edge AI systems. High-Numerical Aperture (NA) EUV lithography is a more advanced version of standard EUV lithography. Both use extreme ultraviolet light to print chip features, but High-NA EUV uses a stronger optical setup that can print smaller features more precisely. • In practice, "normal" EUV usually means today's standard 0.33 NA systems, while High-NA EUV uses a 0.55 NA system.
Fabless Manufacturing	A business model in which a company designs a chip but does not make it in its own factory. Instead, it outsources the production to a specialized manufacturer, such as a foundry (defined below). • Examples of fabless semiconductor companies include Nvidia, AMD, Qualcomm, Broadcom, MediaTek, Marvell, and others.
Fabrication Plant (Fab)	A highly specialized facility where raw silicon is processed into semiconductors. A fab uses advanced manufacturing steps such as photolithography (defined above) to build tiny circuits onto wafers. • Intel operates six fabs in Chandler, AZ, including two new facilities, Fab 52 and Fab 62, built to support Intel's advanced logic chip production. Fab 52 was completed and become fully operational in October 2025, and Fab 62 is expected to be production-ready around 2028.
Foundry	A company that takes chip designs from a customer and manufactures them into physical silicon wafers. Foundries often work with "fabless" companies that design chips but do not own their own manufacturing plants. • All foundries use fabs, but not every fab is a foundry. A company can own a fab to make its own chips, but a foundry makes chips on contract for outside customers. • Taiwan-based TSMC is the world's largest foundry, manufacturing chips for companies such as Nvidia, AMD, Apple, Qualcomm, and more.
Graphics Processing Unit (GPU)	A type of computer chip that is widely used to train and run AI models. Unlike a traditional computer processor (CPU), a GPU can perform many calculations at the same time, allowing it to process the large amounts of data needed for AI much more quickly and efficiently.
High-Bandwidth Memory (HBM)	An innovation in memory designed to move data very quickly between memory and a processor by using 3-D chip stacking (see definition above) to place memory very close to the processor. Compared with conventional memory technologies such as DDR and GDDR, HBM provides substantially higher memory bandwidth—often several times greater—while using less power and occupying less physical space. This allows GPUs and other AI chips to operate at full capacity instead of waiting for data.
Tensor Processing Unit (TPU)	A chip produced primarily for mathematical operations used in AI systems, especially large matrix operations. Unlike a general-purpose CPU, a TPU is built

	for narrower tasks, which can make it faster and more efficient for training or running AI models. • TPUs were developed by Google and first deployed internally in 2015. According to Google, they are built for AI workloads such as agentic AI (defined above), code generation, large language models (defined below), media content generation, synthetic speech, recommendation engines, and more.
Wafer	A thin, circular slice of semiconductor materials—usually silicon—that serves as the foundation for manufacturing computer chips. Many individual chips are built on a single wafer before being cut apart and packaged for use in electronic devices.

AI Models and Training

Term	Definition
Adversarial Distillation Attack	A term coined by leading U.S. AI labs that refers to attempts to copy or extract the knowledge and/or behavior of an AI model by repeatedly querying it and using its outputs to train a different (often competing) model. • Anthropic and OpenAI claim that Chinese AI labs have engaged in large-scaled distillation attacks on their frontier models, using fraudulent accounts and proxy services to transfer knowledge from more powerful models. While knowledge distillation (see definition below) is a commonly used technique, these attacks raise security concerns. A USCC April 2026 China Bulletin noted that Chinese AI labs' adversarial distillation attacks allow them to bypass safeguards and effectively circumvent U.S. export controls on advanced semiconductors.
Agentic AI	An AI system that can independently plan, make decisions, and carry out a series of actions to complete complex tasks with little human involvement. It can break a task into smaller steps, remember relevant information, and adjust its actions as needed to achieve a specific goal. Some AI systems use a single agent, while others coordinate multiple AI agents working together. An AI agent is a system or program that is capable of nearly autonomously performing specific tasks on behalf of a user or another system. • The USCC February 2026 China Bulletin noted that China scrutinized Meta's acquisition of Manus, a Singapore-based AI firm with Chinese roots whose AI agent gained massive popularity in early 2025.
AI Diffusion	Large-scale integration and adoption of AI across existing systems, business operations, and organizations with the goal of improving efficiency, productivity, and innovation.
AI Hallucinations	Incorrect, incoherent, or misleading results generated by an AI model, often caused by a variety of factors such as insufficient training data (defined below), incorrect assumptions made by the model that rely on pattern prediction, or biases in the data used to train the model. AI hallucinations are harmful because they can give users false information that sounds correct, which can lead to mistakes, poor decisions, and misplaced trust

	in the AI system. The risk is especially serious when AI is used in high-stakes settings such as healthcare, law, finance, customer support, or public policy.
AI Inference	The process of using a trained model to generate predictions and perform tasks based on new data beyond those used during the original training phase. To put it simply, while trained models learn more new skills, inference models apply the skill to do a job.
AI Slop	Low-quality AI-generated content, often mass-produced with limited concern for accuracy or value. The term can apply to text, images, audio, video, memes, or spam-like posts.
AI Training	The process by which a model learns how to perform tasks by exposing it to massive, diverse datasets such as text, image, audio, or video and teaching it to recognize patterns and make predictions. AI training is often the most compute-intensive process of building an AI model, especially for frontier models.
Application Program Interface (API)	A software interface that allows different computer programs to exchange data and services, acting as a messenger without knowing the internal functionality of the programs. For example, a company can use an API to connect its website or software to a chatbot model, allowing users to ask questions and receive AI-generated answers inside that service. • When a user asks ChatGPT a question in an app or website, the software sends that question through an API to the underlying AI model, which processes it and returns an answer through the API.
Base Models	Large AI models that have been trained on broad datasets but have not yet been customized for specific tasks or modified with additional safety features. They often serve as the starting point for creating more specialized AI systems, such as chatbots, coding assistants, or image-generation tools. Base models are also often referred to as foundation models.
Benchmarks	Standardized tests used to measure and compare how well AI models perform on specific tasks such as answering questions, writing code, recognizing images, or following instructions. • Although useful to an extent, AI benchmarks are often criticized for being misleading because a "high score" does not always mean a model will perform equally well in real-world use. They can often overstate model performance, especially when the tests are narrow, flawed, or easily gamed. • There is currently no single global standard for benchmarking AI, although multiple international efforts are underway to develop more consistent evaluation methods.
Closed (or Proprietary) Model	A model that is accessible only through the developer's interface (e.g., ChatGPT) or an approved API. Closed models are often juxtaposed with open models (see definition below) in which developers often publish the model weights and/or allow free downloads and customization.
Derivative Models	AI models that are modified from base models to better perform specific functions, such as customer service or coding assistance. Some examples of derivative models include Stanford's Alpaca, Large Model Systems' (LMSYS) Vicuna, and University of California Berkeley's Koala, all of which were derived

	from Meta's LLaMa model. Another example, Salesforce's Einstein GPT, was derived from OpenAI's GPT-3 and GPT-4 and Salesforce's existing Einstein AI platform to improve customer service for both internal employees and external clients. • A 2026 USCC staff paper noted that by the end of 2025, Alibaba's Qwen family had over 100,000 derivatives on Hugging Face, a global open model collaboration platform.
Embodied AI	AI that can perceive, interact with, and act in the physical world through machines such as robots, vehicles, or drones. Unlike AI systems that operate only in software (such as ChatGPT), embodied AI uses sensors to gather information about its surroundings and can take physical actions in response. Examples of embodied AI include self-driving cars, delivery drones, autonomous delivery robots, and surgical robots assisted by AI. • A 2026 USCC staff paper noted that Chinese scientists view embodied AI as a potential pathway to highly capable AI such as artificial general intelligence. • A 2024 USCC staff issue brief highlighted the Chinese government's efforts to develop and deploy humanoid robots—another example of embodied AI—across key industries such as manufacturing, agriculture, healthcare, military, and law enforcement.
Finetuning	The technical process of adapting a base AI model for specific use cases such as customer service, legal document review, financial forecasting, and more. For example, BloombergGPT is a large language model finetuned on financial data to perform tasks associated with financial analysis. • A 2026 USCC report noted that the most downloaded model on Hugging Face in late 2025 was not a frontier LLM but rather a small, specialized video captioning model—ByteDance's Tarsier2-Recap-7b, finetuned from an Alibaba base model known as Qwen2-VL-7B-Instruct.
Frontier Models	The most capable AI models available at a given time, achieving top performance across complex tasks like coding, math, and processing multimodal (see definition below) data (e.g., text, images, and audio), often trained with the largest datasets and most compute available.
Knowledge Distillation	A machine learning technique to transfer learned knowledge from a "teacher" AI model to a "student" AI model. The technique is not new, having first emerged as a concept in a 2006 academic paper, with the term officially being coined in 2015. Knowledge distilled could include outputs (i.e., final predictions such as image classification), intermediate layers (i.e., step-by-step problem-solving process), and relationships (i.e., structural layers between different data samples). Knowledge distillation aims to reduce the amount of memory and compute needed to train a model by transferring specialized, diverse, multi-task knowledge from the "teacher" to the "student" model, thereby negating the need for the "student" model to go through the same rigorous amount of training. The "student" model (if left as is), however, usually remains less powerful than the "teacher" model.

Large Language Model (LLM)	An AI model that generates text when queried by predicting the next token (see definition below) based on patterns learned from large datasets. This is the type of model most consumers associate with AI. During training of an LLM, input text is split into tokens (see definition below) and encoded so the model can learn and predict possible continuations.
Mixture of Experts	An AI model design that uses only the parts of the model most relevant to a given task or input rather than using the entire model every time. This approach can improve speed and efficiency by reducing the amount of computing power needed while maintaining strong performance. • The USCC March 2026 China Bulletin noted that Alibaba's Qwen 3.5 model scales 128 experts in its predecessor to 512.
Model Weights	A subset of parameters (see definition below) in a neural network—a brain-inspired technique that teaches models to process data—that determines how models process information, predicts how strongly different inputs influence outputs, and generates outputs.
Multimodal	The capacity of an AI model to process and understand multiple types of data ("modalities")—for example, text, still images, audio, and video. A multimodal model can handle more than one type of data on the input or output side or both.
Open Model	A model that publicly releases its weights (see definition above). Often referred to as open-weight models, which release only parameters (defined below). Open-weight models also have varying degrees of licensing permissions, which may allow developers to independently understand, audit, recreate, and build upon the model, unlike closed or proprietary models (see definition above). • A 2026 USCC report found that China is betting on open models to advance its AI capabilities and increase adoption across different sectors and is currently leading the world in such models. Examples include China's DeepSeek V4 and the United States' OpenAI's gpt-oss.
Open-Source Model	A model whose key components—including its code, weights, and training data—are made publicly available so others can use, study, modify, and share it. • Stability's Stable Diffusion 2 model is an open-source text-to-image AI model that was released in 2022 and included information on the model's weights, training data, architecture, and more.
Parameters	The learned numerical values inside an AI model that determine how it processes inputs (i.e., training data) and generates outputs. Parameters include model weights (see definition above), biases that help models generalize, and other trainable values.
Reinforcement Learning from Human Feedback (RLHF)	A method of training in which human feedback is used to help an AI model produce better and more preferred outputs. • ChatGPT is one notable application of RLHF, where users responded to or rewarded the model with good or bad reactions to generated outputs. While RLHF can enhance output accuracy, it also tends to be more expensive and time-consuming.
Small Language Model	Language models that are trained on fewer parameters and smaller datasets for specialized tasks, requiring less compute and data than LLMs (see

	definition above). Examples of small language models include Meta's LLaMa 3.2-1B, Microsoft's Phi-3.5-Mini-3.8B, and Google's Gemma3-4B.
Software Programming Platforms	The software used for running computing tasks across several AI chips (usually GPUs) at the same time. These platforms help the AI chips handle many calculations in parallel, thereby increasing their speed and efficiency in handling any given task. • Nvidia's Compute Unified Device Architecture (CUDA) has become the de facto standard for AI training and inference, enabling developers to harness thousands of chips in parallel.
Tokens	Units of text that AI language models use to process language. A token can be a whole word, part of a word, punctuation mark, or other sequence of characters. AI models read, generate, and count text in tokens rather than words. For example, a word such as "unbelievable" may be broken into smaller components such as "un," "believe," and "able." AI models generate text by predicting likely next tokens in a sequence. Tokens are also commercially and operationally important because AI firms often price services based on token usage, while higher token volumes increase computing demand, energy consumption, and the need for advanced chips and data center capacity.

Data

Term	Definition
Data Curation	The process of collecting, cleaning, selecting, labeling, organizing, and maintaining data so it can be used effectively to train, test, or run AI systems. • Good data curation helps improve model quality because AI models depend heavily on the quality of data they learn from or use. Poorly curated data can make models less accurate, less reliable, or more biased.
Data Labeling	The process of adding tags or categories to raw data so an AI model can learn what the data means. Without labeled examples, or if the labels are inconsistent, an AI model's outputs can be inaccurate. • For instance, a person may label something "positive" or "negative" to identify when an X-ray shows a tumor. Once this data is then used to train an AI model, the model will have a better chance of identifying tumors in X-rays.
Data Poisoning	The deliberate corruption of training data to cause an AI system to perform poorly or behave incorrectly. It might involve adding false data at the training stage, changing existing data at the inference stage, or biasing the dataset so the model becomes less accurate, less reliable, or easier to exploit. Data poisoning can cause a model to make incorrect predictions, reveal hidden cybersecurity vulnerabilities, or produce manipulated outcomes. Examples of data poisoning include: • Training a model on misleading medical data, which could lead to incorrect diagnoses and treatment recommendations. • Manipulating content moderation systems on social media platforms to spread false information.

Data Provenance	The record that shows the source and history of the data used by a system. It includes information about where the data came from, how it was created, and how it has been changed over time. For AI, data provenance matters because it helps people judge whether data is trustworthy, authentic, and appropriate to use for training a model.
The "Data Wall"	The point at which AI models stop improving much because the available training data (defined below) has been largely used up or no longer adds enough value. To date, AI progress has depended heavily on large amounts of high-quality training data. If that data becomes scarcer, future model improvement could slow unless developers use synthetic data (defined below).
Raw Data	Information in its original, unprocessed form before it has been cleaned, labeled, organized, or analyzed for use in an AI system. Raw data can include text, images, audio, video, sensor readings, web pages, spreadsheets, or other source material collected from the real world. Raw data is the foundation for training and evaluating AI models. However, raw data by itself is usually not ready for use, so it often needs labeling (defined above) and curation (defined above) before a model can learn from it effectively.
Synthetic Data	Artificially generated data from AI models and algorithms that is primarily used to train and improve AI systems at scale. The data mimics real-world data so that it can supplement or replace real datasets that are proprietary, difficult to obtain, or becoming finite. The advantages of synthetic data are that it is scalable, cheaper than collecting real-world data, preserves privacy, and can be tailored to specific needs. However, the data often lacks the complexity of real-world data, can amplify existing biases, and risks degrading model performance over time if AI systems are trained repeatedly on AI-generated data.
Training Data	The information used to teach an AI model how to recognize patterns, make predictions, or generate outputs. It can include text, images, audio, video, spreadsheets, or other examples the model learns from during training. The quality, quantity, and diversity of training data strongly affect how well an AI system performs.

Ecosystem

Term	Definition
AI Ecosystem	Encompasses all elements involved in developing, deploying, and using AI, including technologies, data, talent, governance, funding, and infrastructure (i.e., data centers).
AI Infrastructure	The hardware and software needed to develop, run, and manage AI systems. This includes computing chips, data centers, networking equipment, and the software tools that support AI applications and tasks.

	AI infrastructure is the underlying foundation that powers the AI tech stack (defined below).
AI Tech Stack	The collection of technologies built on top of one another that are necessary to build, train, deploy, and run AI systems. The AI tech stack generally encompasses three interdependent layers: infrastructure, models, and applications.
Data Centers	Large facilities housing thousands of specialized chips, servers, and networking systems designed to process, store, and distribute data and handle the massive computing demands of training and running AI models.
Hyperscaler	Very large cloud-computing companies that operate massive data centers and provide computing, storage, and networking services at enormous scale. They are built to expand capacity quickly so customers can run large digital services, AI workloads, and other computing tasks without needing to build all the infrastructure themselves. • Examples of hyperscalers include Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Alibaba Cloud, and more.
Server Racks	The physical structures (usually metal frames or cabinets) in data centers that house high-performance servers, chips, cables, and cooling equipment. • An AI chip (i.e., a GPU) is usually not useful by itself in a data center because it needs a host server with CPUs, memory, storage, an operating system, and expansion slots to run AI workloads properly. That is why vendors often package GPUs inside GPU server racks rather than treating them as standalone products for enterprise AI buyers such as hyperscalers (defined above).