AI News

Glossary

2026-05-02 - 14:29

Compiled with Claude AI

Core AI concepts

Foundational vocabulary

A

Agent / AI Agent *(also: AI Agent, AI agent, agent)*

An AI system that can autonomously plan, make decisions, use tools, and execute multi-step tasks with minimal human intervention. Example: Claude Code spawning sub-agents to write, test, and deploy code.

Agentic AI *(also: agentic, agentic workflow, agentic systems)*

Describes AI systems that plan, decide, and execute multi-step tasks with limited human oversight, rather than responding to a single prompt at a time. The shift from chatbots to agentic systems is one of the defining trends of 2026.

AGI (Artificial General Intelligence) *(also: AGI, Artificial General Intelligence)*

A hypothetical AI system that can understand, learn, and apply knowledge across any intellectual task at a human level or beyond. No AGI exists yet; current AI is considered "narrow" (good at specific tasks).

Alignment

The challenge of ensuring AI systems behave in accordance with human values and intentions. A core research area at labs like Anthropic, focused on making AI safe and trustworthy.

API (Application Programming Interface) *(also: API)*

A set of rules that allows software applications to communicate with each other. AI APIs let developers integrate models like GPT or Claude into their own apps and workflows.

Autoregressive Model

A model that generates output one piece (token) at a time, with each new piece depending on all the previous ones. Most current LLMs (GPT, Claude, Gemini) work this way.

C

Chain-of-Thought *(also: CoT, Chain-of-Thought prompting)*

A prompting technique where the AI is encouraged to think step by step, breaking complex problems into smaller reasoning steps before arriving at an answer. Improves accuracy on reasoning tasks.

Chatbot

An AI application designed for conversational interaction with users via text or voice. Examples: ChatGPT, Claude, Gemini. Modern chatbots are powered by LLMs.

Context Engineering

The discipline of designing what information an AI agent or model receives, when it receives it, and in what format — as distinct from prompt engineering, which focuses on how the instruction is phrased. In multi-agent systems, context engineering determines which tools, memory, and intermediate results each agent sees, directly affecting reliability and output quality. Emerging as the defining skill of agentic AI in 2026.

Context Window *(also: context window)*

The maximum amount of text (measured in tokens) an AI model can process in a single conversation. Larger windows allow the AI to process more of the conversation or read longer documents. Example: Gemini 3.1 Pro has a 2 million token context window.

D

Deepfake

AI-generated media (video, audio, images) that realistically depicts people saying or doing things they never actually did. Raises significant ethical and misinformation concerns.

F

Few-Shot Learning

An AI's ability to perform a task after being shown just a few examples, rather than requiring thousands of training examples. A key strength of modern LLMs.

Foundation Model *(also: foundation models)*

A large-scale model trained on broad data, designed to be adapted to many downstream tasks. The term overlaps with frontier model in industry coverage; foundation model is the more general descriptor, frontier model emphasises being at the leading edge of capability.

G

Generative AI *(also: GenAI, Generative Artificial Intelligence)*

AI systems that can create new content — text, images, music, video, code — rather than just analysing or classifying existing data. The umbrella term for tools like ChatGPT, Midjourney, and Suno.

Grounding

Connecting an AI model's responses to specific, verifiable sources of information (documents, databases, web results) to reduce hallucinations and improve accuracy.

Guardrails

Safety mechanisms built into AI systems to prevent harmful, biased, or inappropriate outputs. Can include content filters, usage policies, and constitutional AI approaches.

H

Hallucination

When an AI generates information that sounds confident and plausible but is factually incorrect or entirely fabricated. A known limitation of all current LLMs, though hallucination rates are falling significantly across frontier models in 2026.

I

Inference

The process of an AI model generating an output (answer, image, etc.) from an input (prompt). Distinguished from "training," which is how the model learns. What happens every time you ask Claude a question.

P

Prompt Engineering

The skill of crafting effective prompts to get the best results from AI models. Includes techniques like providing context, giving examples, specifying format, and asking for step-by-step reasoning. Increasingly supplemented by context engineering in agentic systems.

T

Tool Use *(also: function calling, tool calling, function call)*

The capability of an AI model to call external functions, APIs, or tools as part of its response. The model decides when to invoke a tool, what arguments to pass, and how to integrate the result. Function calling is the technical foundation of agentic systems.

Z

Zero-Shot Learning

An AI's ability to perform a task it has never been explicitly trained on, without any examples. Modern LLMs can often handle novel tasks purely from their general training.

Models and architectures

Model families and structural patterns

C

Claude

Anthropic's family of AI models. Named after Claude Shannon, the father of information theory. Comes in three sizes: Haiku (small/fast), Sonnet (balanced), and Opus (most capable).

Constitutional AI

An approach developed by Anthropic where the AI is trained to follow a set of principles (a "constitution") to be helpful, harmless, and honest, reducing reliance on human feedback for safety.

D

Diffusion Model

A type of AI model that generates content by progressively removing noise from random input. The technology behind most image generators (Midjourney, DALL-E, Stable Diffusion) and emerging video models.

F

Frontier Model

The most capable and advanced AI models available at any given time. Currently includes GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro, and similar top-tier models.

G

GPT (Generative Pre-trained Transformer) *(also: GPT)*

OpenAI's family of language models. "Generative" means it creates content, "Pre-trained" means it learned from vast amounts of data, and "Transformer" is the neural network architecture it uses.

L

LLM (Large Language Model) *(also: LLM, Large Language Model)*

An AI model trained on massive amounts of text data that can understand and generate human language. The technology behind ChatGPT, Claude, Gemini, and similar tools. "Large" refers to billions of parameters.

M

Mixture of Experts *(also: MoE)*

A model architecture where only a subset of the model's parameters are activated for each input, making it faster and cheaper to run. Used by Mistral Small 4 (119B total params, just 6.5B active per query) and Meta Llama 4 Maverick.

Multimodal

An AI system that can process and/or generate multiple types of media: text, images, audio, video. Most frontier models in 2026 are multimodal. Example: you can show Claude an image and ask about it.

N

Neural Network

A computing system inspired by the human brain, made up of layers of interconnected nodes that process information. The foundation of modern AI, including LLMs and image generators.

O

Open Source / Open-Weight *(also: open-weight, open source AI)*

AI models whose code and/or trained weights are publicly available for anyone to download, use, modify, and host. Examples: Llama 4 (Meta), Mistral Small 4, Stable Diffusion, Flux 2. "Open-weight" specifically means the trained model is available but the training code or data may not be.

R

Reasoning Model

An AI model specifically designed to solve complex problems through extended, step-by-step logical reasoning, often thinking for longer before answering. Examples: OpenAI o-series, DeepSeek-R1. They trade speed for accuracy.

T

Transformer

The neural network architecture that powers most modern AI models (GPT, Claude, Gemini, Llama). Introduced by Google in 2017, it excels at understanding relationships between words across long texts.

Capabilities and benchmarks

Evaluation frameworks and capability metrics

A

AGI Benchmark *(also: ARC-AGI-2)*

ARC-AGI-2 is an independent benchmark designed to resist gaming by AI models, testing abstract pattern reasoning rather than memorised knowledge. Gemini 3.1 Pro scored 77.1% in March 2026, more than doubling its predecessor's score.

ARC-AGI *(also: ARC-AGI-2)*

Abstraction and Reasoning Corpus, a benchmark designed to measure abstract pattern recognition. ARC-AGI-2 is the harder second-generation version released in 2025. Treated as a proxy for general-reasoning capability and the hardest published benchmark for current frontier models.

Artificial Analysis Intelligence Index *(also: Intelligence Index, AA Intelligence Index)*

A composite reasoning score maintained by independent evaluation firm Artificial Analysis, combining GDPval-AA, GPQA Diamond, Tau2-bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity’s Last Exam, and CritPt. Index v4.0 was the active methodology through Q1 2026, with Gemini 3.1 Pro and GPT-5.4 effectively tied at the top at 57 points.

B

Benchmark

A standardised test used to measure and compare AI model performance. Examples: SWE-Bench (coding), MMLU (general knowledge), GPQA (graduate-level science), HumanEval (code generation), OSWorld (computer use), ARC-AGI-2 (abstract reasoning).

BrowseComp

A benchmark that measures multi-step web research capability: browsing, synthesising, and reasoning across multiple pages. GPT-5.5 Pro led at 90.1% in April 2026; Claude Opus 4.7 at 79.3%. The benchmark is one of the few where OpenAI consistently leads Anthropic in the 2026 frontier comparisons.

C

Computer Use

The ability of an AI model to interact directly with a computer's graphical interface — clicking buttons, filling forms, navigating applications, and executing multi-step workflows — without a purpose-built API. GPT-5.4 became the first model to exceed human expert performance on the OSWorld benchmark (March 2026), scoring 75.0% against a human baseline of 72.4%.

CursorBench

An AI coding benchmark produced by the Cursor IDE on real developer workflows rather than synthetic GitHub issues. Treated as a real-workload signal for production coding assistance. Claude Opus 4.7 reached 70% on CursorBench in April 2026, up from 58% on Opus 4.6.

E

Embedding *(also: embeddings)*

A numerical representation of text, images, or other data in a format that AI models can process. Similar concepts get similar numbers, allowing AI to understand relationships between ideas.

F

FrontierMath

A research-grade mathematics benchmark with tiers of increasing difficulty. Tier 4 contains problems at the edge of mathematics research. GPT-5.5 reached 35.4% on FrontierMath Tier 4 in April 2026, with Claude Opus 4.7 at 22.9%.

G

GDPval *(also: GDPval-AA)*

A benchmark that evaluates AI performance on economically valuable professional tasks — financial modelling, legal document analysis, expert-level research, and similar knowledge work. Reported in Elo ranking form. Claude Sonnet 4.6 led GDPval-AA at 1633 Elo as of February 2026, ahead of Claude Opus 4.6 at 1606 and Gemini 3.1 Pro at 1317.

GPQA Diamond *(also: GPQA)*

A benchmark testing graduate-level reasoning in biology, chemistry, and physics. The Diamond subset contains the hardest 198 questions written by domain experts. Frequently cited as a measure of expert-level scientific understanding.

H

HumanEval

A benchmark of 164 hand-written programming problems used to measure code generation capability. One of the earliest standardised coding benchmarks; partly superseded by SWE-bench for agentic code work.

Humanity's Last Exam *(also: HLE)*

A benchmark of expert-grade questions across science and reasoning, designed to remain unsolved by current frontier models for as long as possible. Models are evaluated with and without tool access. Claude Opus 4.7 reached 46.9% without tools and GPT-5.5 reached 57.2% with tools by April 2026.

I

Inference-Time Scaling *(also: test-time compute, inference-time scaling)*

Spending additional compute resources during inference (generation) to improve output quality. The AI thinks harder by exploring more possibilities before responding. A key quality lever in 2026, complementing training-time improvements.

L

Long Context *(also: long-context, long context window)*

The ability of an AI model to process very large input prompts in a single pass — measured in millions of tokens by 2026 standards. A 1-million-token window holds the equivalent of several books or an entire codebase.

M

METR

Model Evaluation and Threat Research, a research organisation that produces task-completion time-horizon estimates for frontier models. The horizon metric measures how long a model can sustain coherent multi-step work before reliability degrades. Claude Opus 4.6 was estimated at 14 hours 30 minutes at the 50% mark in February 2026, the highest reported among frontier models at the time.

MMLU *(also: MMLU-Pro)*

Massive Multitask Language Understanding, a benchmark of 57 academic and professional knowledge tests. MMLU-Pro is a harder revision released to keep up with frontier model performance.

MRCR *(also: MRCR v2)*

Multi-Round Co-reference Resolution, a long-context retrieval benchmark that measures a model's ability to find and reason across information distributed throughout very long inputs. MRCR v2 evaluates retrieval at the 1-million-token mark. Claude Opus 4.6 reached 78.3% on MRCR v2 at 1M tokens, against 36.6% for GPT-5.4 and 18.3% for Gemini 3.1 Pro at the same context length.

N

Natural Language Processing *(also: NLP)*

The branch of AI focused on enabling computers to understand, interpret, and generate human language. LLMs are the latest and most powerful NLP technology.

O

OSWorld *(also: OSWorld-Verified)*

A benchmark that measures autonomous desktop task completion: file management, browser navigation, multi-application workflows. Human-expert baseline is 72.4%. GPT-5.4 became the first model to cross this threshold in March 2026 at 75.0%; Claude Opus 4.7 reached 78.0% in April 2026. OSWorld-Verified is the curated reproducible variant.

R

RAG (Retrieval-Augmented Generation) *(also: RAG, Retrieval-Augmented Generation)*

A technique that enhances AI responses by first retrieving relevant information from external sources (documents, databases), then using that information to generate more accurate, grounded answers.

S

SWE-Bench

A benchmark that tests AI models on their ability to solve real-world software engineering problems from GitHub issues. Gemini 3.1 Pro scores 80.6% on SWE-Bench Verified as of March 2026, among the highest published scores.

T

Terminal-Bench *(also: Terminal-Bench 2.0)*

A benchmark that tests command-line task execution by AI agents: shell scripting, process management, multi-step terminal workflows. Terminal-Bench 2.0, the version active throughout 2026, was led by GPT-5.5 at 82.7% and Claude Opus 4.7 at 69.4% by April 2026.

W

World Model

An AI system that understands physical reality and cause-and-effect relationships, not just language patterns. Distinct from LLMs, which operate in linguistic space. Championed by Yann LeCun.

Creative AI tools

Generative tools for media production

I

Image-to-Video *(also: I2V)*

The process of animating a still image into a video clip using AI. Considered the most controllable approach to AI video generation since you define the starting frame precisely.

Inpainting

An AI technique for editing specific regions of an image or video while keeping the rest intact. Example: selecting a person's shirt in a photo and changing its colour.

L

LoRA (Low-Rank Adaptation) *(also: LoRA)*

A technique for efficiently fine-tuning large AI models by training only a small number of additional parameters rather than the entire model. Popular in the Stable Diffusion community for creating custom styles.

O

Outpainting

An AI technique for extending an image beyond its original borders, generating new content that seamlessly continues the existing scene.

S

Stable Diffusion

An open-source image generation model created by Stability AI. Widely used as the foundation for many image generation tools and custom models. Version 3.5 is the current production standard.

Stem Separation

In AI music, the ability to split a generated track into individual components (vocals, drums, bass, melody) for separate editing and remixing.

T

Text-to-Image *(also: T2I)*

Generating images from written text descriptions (prompts). The core functionality of tools like Midjourney, DALL-E, and Stable Diffusion.

Text-to-Video *(also: T2V)*

Generating video clips from written text descriptions. Offered by Veo 3.1, Runway, Kling, and Seedance. OpenAI discontinued the standalone Sora app in March 2026; the model remains accessible via ChatGPT.

Coding and development

AI-assisted programming concepts

A

Agent Teams

A feature where multiple AI agents coordinate in parallel, each handling a portion of a larger task. Introduced by Anthropic in Claude Opus 4.6 (February 2026).

C

Claude Code

Anthropic's command-line tool that enables developers to delegate coding tasks to Claude directly from their terminal. Can autonomously write, test, and debug code. Available in GitHub Agent HQ from February 2026.

Cowork

Anthropic's desktop tool (released January 2026) that brings Claude Code's agentic capabilities to non-technical users via a graphical interface with local file access, recurring task scheduling, and plugin support.

V

Vibe Coding

A trend where non-programmers use AI coding tools by describing what they want in natural language, letting the AI handle all the actual programming. Popularised by Claude Code's accessibility. Mark Zuckerberg was reported to be vibe-coding Meta's monorepo using Claude Code in March 2026.

Z

Zero-Day Vulnerability *(also: zero-day)*

A previously unknown security flaw in software that attackers could exploit before developers are aware of it. Claude Opus 4.6 found 500+ of these in open-source code during testing (February 2026).

Infrastructure and protocols

Technical foundations of model deployment

H

Hyperscaler *(also: hyperscalers)*

A cloud provider operating data centres at very large scale — primarily Amazon Web Services, Microsoft Azure, Google Cloud, and Oracle Cloud. Hyperscalers are central to AI deployment because frontier model training and inference require dedicated compute capacity at scale.

K

KV Cache *(also: key-value cache)*

Key-Value cache. A memory structure that stores intermediate computations during transformer inference, allowing the model to avoid recomputing them on each token. KV cache size grows with context length and is a primary memory bottleneck for long-context inference.

M

MCP (Model Context Protocol) *(also: MCP, Model Context Protocol)*

An open standard managed by the Linux Foundation for how AI agents connect to and use external tools and data sources. Crossed 97 million monthly SDK downloads in March 2026 with over 5,800 community and enterprise servers. Think of it as a universal plug for AI tool integration.

P

Parameters

The internal numerical values that an AI model learns during training. More parameters generally means more capability. GPT-5.4 and Claude Opus 4.6 have hundreds of billions of parameters.

Q

Quantization *(also: quantize, quantized)*

A compression technique that reduces the numerical precision of a model's weights to make it smaller and faster, with limited loss of accuracy. Modern methods can quantize from 16-bit down to 4-bit or lower while preserving most of the model's capability.

S

Speculative Decoding

An inference optimisation in which a smaller, faster model proposes several candidate tokens which the main model then verifies in parallel. The result is faster generation with no loss in output quality.

T

Token *(also: tokens)*

The basic unit of text that AI models process. Roughly equivalent to three-quarters of a word in English. A 1 million token context window can hold approximately 750,000 words — about 10 to 12 novels.

TPU (Tensor Processing Unit) *(also: TPU, TPUs)*

A custom chip designed by Google for AI workloads. TPUs are deployed widely in Google Cloud and used to train and serve Gemini models. Distinct from GPUs (general-purpose graphics processors adapted for AI) and from custom AI silicon developed by other vendors.

Training

The process of teaching an AI model by feeding it large amounts of data so it can learn patterns. Pre-training uses general data; fine-tuning uses specialised data. Training frontier models costs tens to hundreds of millions of dollars.

Training techniques

Methods for teaching AI models

D

Distillation *(also: model distillation)*

A training technique where a smaller "student" model is trained to replicate the behaviour of a larger "teacher" model, rather than learning directly from raw data. The result is a compact model that retains much of the teacher's capability at a fraction of the inference cost. Example: many efficient open-weight models use distillation from frontier models to achieve near-frontier performance at significantly lower cost.

DPO (Direct Preference Optimization) *(also: DPO, Direct Preference Optimization)*

A training method that aligns an AI model with human preferences directly from preference data, without the explicit reward model required by RLHF. Often used as a simpler, more stable alternative.

E

Embedding Training *(also: vector training)*

The process of training a model to produce numerical representations (embeddings) that capture the semantic meaning of text or other data. These embeddings are used in RAG systems to retrieve relevant content efficiently.

F

Fine-Tuning *(also: fine-tuning, finetuning)*

The process of further training a pre-trained AI model on a specific dataset to specialise it for a particular task, domain, or style. Like giving a general expert specialised training.

P

Prompt *(also: prompts)*

The text instruction or question you give to an AI model. The quality and specificity of your prompt significantly affects the quality of the output.

R

RLHF (Reinforcement Learning from Human Feedback) *(also: RLHF)*

A training technique where human reviewers rate AI outputs, and the model learns to produce responses that humans prefer. A key step in making LLMs helpful and safe.

Governance and safety

Policy frameworks and safety mechanisms

A

AI Safety Levels (ASL) *(also: ASL, AI Safety Level, ASL-4)*

Anthropic's internal classification of AI capability levels and the corresponding safety measures required for deployment. Higher levels apply progressively stricter testing and access controls. ASL-4 was triggered for the first time in early 2026 with the Claude Mythos model, contributing to the decision not to release it publicly.

D

Digital Omnibus on AI *(also: Digital Omnibus)*

A European Commission proposal adopted in November 2025 to amend the EU AI Act and related legislation. Reached political agreement on March 11, 2026. The proposal extends certain high-risk system obligations under Annex III of the EU AI Act to 2 December 2027, while preserving the original 2 August 2026 schedule for Article 50 transparency obligations on AI-generated content.

E

EU AI Act *(also: AI Act)*

European Union Regulation 2024/1689, the first comprehensive legal framework on artificial intelligence. Adopted in 2024 and phased in through 2027. The 2 August 2026 deadline marks application of most obligations, including transparency rules for AI-generated content and requirements for general-purpose AI models.

G

GPAI (General Purpose AI) *(also: GPAI, General Purpose AI)*

Regulatory category in the EU AI Act covering AI models with broad capabilities that can perform a wide range of tasks — typically large language models. GPAI providers must publish summaries of training data and meet transparency requirements.

P

Project Glasswing *(also: Glasswing)*

Anthropic's research and access programme through which Claude Mythos and other capability-restricted models are made available to a limited set of vetted enterprise partners. Announced April 7, 2026, alongside the confirmation of Claude Mythos. The programme includes safeguards on cybersecurity-relevant uses and a Cyber Verification Program for legitimate offensive-security research.

T

Texas TRAIGA *(also: TRAIGA, Texas Responsible Artificial Intelligence Governance Act)*

The Texas Responsible Artificial Intelligence Governance Act, effective 1 January 2026. Establishes disclosure requirements for AI system deployers in Texas, prohibits certain discriminatory uses of AI in employment and housing decisions, and creates state-level enforcement authority. The second comprehensive US state AI law to take effect, after the Colorado AI Act (June 2026).