https://www.nytimes.com/2026/11/18/science/ai-artificial-intelligence-terminology-glossary.html

—
give me a list of the top 20 jargon terms of art in AI -- like rack, alignment, and sandbox.
—

Here is a curated list of top terms of art and industry jargon spanning AI safety, infrastructure, training, engineering, and deployment:

1. Alignment
The process of steering an AI model's behavior to match human intent, values, and ethical principles. This involves techniques like RLHF (Reinforcement Learning from Human Feedback) to prevent harmful or off-topic outputs.

2. Sandbox
A restricted, isolated environment where unverified AI agents, executable code, or experimental models can run safely without risk of corrupting external systems, leaking data, or accessing network resources.

3. Compute Rack (or Rack)
The core unit of hardware infrastructure powering AI. A rack physically houses tightly coupled server nodes fitted with high-performance GPUs or TPUs, liquid cooling, and ultra-high-speed interconnects designed for heavy training and inference workloads.

4. Hallucination
When a generative model confidently produces outputs that are factually incorrect, nonsensical, or entirely fabricated, despite appearing plausible.

5. Context Window
The maximum volume of input and memory tokens (text, code, image frames) an LLM can process, retain, and reason over in a single prompt or conversation turn.

6. RAG (Retrieval-Augmented Generation)
An architecture that enhances LLM responses by fetching relevant, real-time factual documents from an external vector database before generating an answer.

7. Agentic AI
AI systems engineered not just to generate text or images, but to autonomously plan multi-step workflows, make decisions, invoke tools/APIs, and execute complex goals with minimal human intervention.

8. Vibe Coding
A software engineering term describing a practice where a developer builds applications primarily by describing desired features and behavior in natural language to an AI coding assistant, letting the model write the underlying code.

9. RLHF (Reinforcement Learning from Human Feedback)
A fine-tuning method that uses human evaluations to train a reward model, which then shapes the core AI model to generate safer and more helpful outputs through reinforcement learning.

10. Catastrophic Forgetting
The tendency of a neural network to completely lose or overwrite previously learned information when trained on a new, unrelated dataset or task.

11. Quantization
A model compression technique that reduces precision (e.g., converting 16-bit floating-point weights down to 8-bit or 4-bit integers) to shrink memory requirements and speed up inference with minimal performance loss.

12. Distillation
The process of training a smaller, faster "student" model to mimic the outputs and reasoning of a massive, complex "teacher" model, enabling efficient deployment on lightweight hardware.

13. System Prompt
The foundational instruction set given to a model prior to user interactions that defines its core identity, tone, operational boundaries, and rules of engagement.

14. Chain-of-Thought (CoT)
A prompting technique or model capability that forces an AI to explicitly spell out its intermediate reasoning steps before arriving at a final answer, improving accuracy on complex logic and math tasks.

15. Temperature
A hyperparameter that controls the randomness and variability of model outputs. Lower temperatures yield deterministic and focused answers, while higher temperatures yield more creative or varied generation.

16. Latent Space
A high-dimensional mathematical space where an AI model maps compressed representations of data (text, concepts, or images), allowing semantically similar ideas to sit close to one another.

17. Multimodal
A model architecture capable of simultaneously processing, understanding, and generating multiple data formats—such as text, vision, audio, and code—within a unified framework.

18. Model Context Protocol (MCP)
An open standard that provides a universal format for connecting AI models to external enterprise data sources, development environments, and software tools without custom API integrations.

19. Mechanistic Interpretability
A subfield of AI safety focused on reverse-engineering neural networks to inspect individual neurons and weights, aiming to understand precisely *how* models make decisions internally rather than treating them as black boxes.

20. Token
The foundational unit of data processed by language models. Tokens can represent single characters, sub-words, or complete words depending on the tokenization scheme.