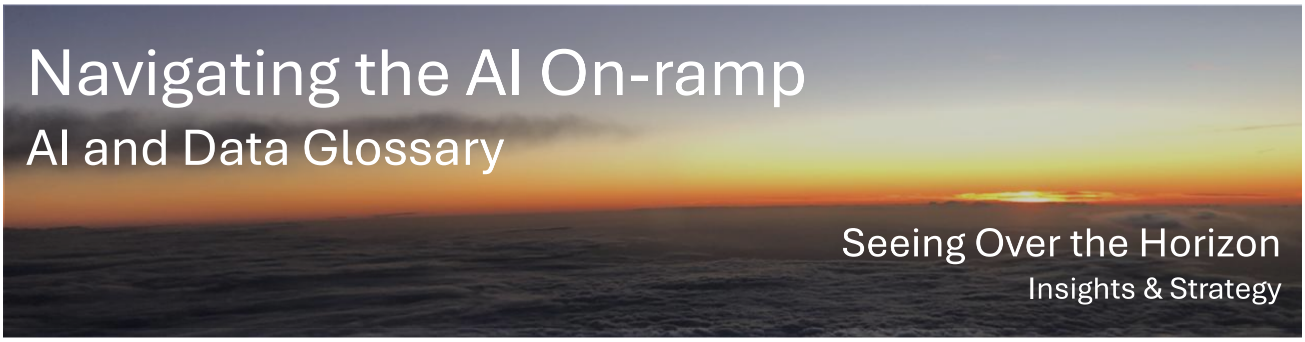


Navigating the AI On-ramp

AI and Data Glossary

Seeing Over the Horizon

Insights & Strategy

This report

This glossary provides brief explanations of some of the sector-specific phrases and jargon used in the reports in the series of studies looking at what lies ahead as AI penetrates through business and elsewhere.

Navigating the AI On-Ramp : https://www.peterosborn.com

Glossary

Glossary

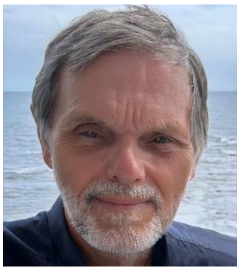
- **Adapter**: Small neural network modules that are inserted into pre-trained foundation models to adapt them to new tasks or domains without requiring full model retraining.
- **Agentic AI**: AI systems that function as autonomous agents capable of making decisions, learning from interactions, adapting to changing environments, and taking independent action to achieve specific goals.
- **AI Data Paradox**: The challenge where increasing data volume doesn't automatically improve AI model performance, as data quality, relevance, and proper utilization often matter more than quantity alone.
- **AI temperature**: A parameter that controls the randomness of text generation in large language models by adjusting the probability distribution during token selection, where higher values increase creativity and lower values make outputs more deterministic.
- **Apache Iceberg**: An open-source high-performance table format for huge analytic datasets that brings the reliability and simplicity of SQL tables to big data environments, enabling modern lakehouse architectures.
- **Autophagous loops**: Self-consuming feedback loops in AI systems where the model's own outputs become part of its training data, potentially leading to degradation or bias amplification over time.
- **Backward Chaining**: A goal-driven inference technique in AI that starts with the goal and works backward through rules to find known facts that support the goal, commonly used in expert systems and diagnostic applications.

- **Burstiness signal**: A metric used to identify AI-generated text by measuring the variation in sentence length and structure, where human writing typically shows more variation (high burstiness) compared to AI-generated content.
- **Chain of thought**: A prompt engineering technique that improves AI reasoning by instructing models to generate step-by-step explanations of their reasoning process before arriving at a final answer.
- **Data lake**: A centralized repository that stores vast volumes of structured, semi-structured, and unstructured data in its native, raw format, providing scalable storage for diverse data types without requiring upfront transformation.
- **Data mesh**: A decentralized data architecture approach that treats data as a product and emphasizes domain-oriented ownership, based on four principles: domain ownership, data as a product, self-serve platforms, and federated governance.
- **Data Sovereignty**: The concept that data is subject to the laws and governance structures of the nation or jurisdiction where it is stored or processed, ensuring local control over data privacy and security.
- **Diffusion models**: Generative AI models that create new content by learning to reverse a gradual noising process, starting with random noise and iteratively refining it to produce high-quality images, text, or other data.
- **Dimensionality Reduction Techniques**: Methods that reduce the number of input variables or features in datasets while preserving essential information, including techniques like PCA, LDA, t-SNE, and UMAP to improve computational efficiency and model performance.
- **edge**: Computing infrastructure deployed close to where data is generated and consumed, reducing latency and bandwidth requirements by processing data locally rather than in distant cloud data centres.
- **Egress fees**: Charges imposed by cloud providers for transferring data out of their services or platforms, often representing a significant cost factor in cloud computing and data architecture decisions.
- **Federated decision making**: A governance approach where decision-making authority is distributed across multiple autonomous units or domains while maintaining coordination through shared principles and standards.
- **FLAP-D Markets**: The major European data centre markets of Frankfurt, London, Amsterdam, Paris, and Dublin, representing the primary hubs for colocation and cloud infrastructure in Europe.

- **Foundation models or Large X models**: Large-scale AI models trained on vast datasets that serve as the foundation for various downstream tasks, including large language models (LLMs), large vision models, and multimodal models.
- **Fully sharded data parallel**: A distributed training technique that partitions model parameters, gradients, and optimizer states across multiple devices to enable training of extremely large models that exceed single-device memory capacity.
- **Generative Adversarial Networks**: Machine learning architectures consisting of two neural networks competing against each other - a generator that creates synthetic data and a discriminator that tries to distinguish real from fake data.
- **GPTzero**: An AI detection tool designed to identify AI-generated content by analysing text characteristics such as perplexity and burstiness to distinguish between human-written and machine-generated text.
- **high-density DC**: Data centres designed to support high power density computing workloads, typically accommodating 50-100+ kW per rack to support AI and high-performance computing applications.
- **Homomorphic encryption**: A form of encryption that allows computations to be performed on encrypted data without first decrypting it, enabling secure processing of sensitive data while maintaining privacy.
- **homoglyph substitutions**: A security attack technique that uses visually similar characters from different character sets to create deceptive URLs, domain names, or text, exploiting the human eye's inability to distinguish between similar-looking characters.
- **Hyperscaler**: Large cloud service providers like Amazon, Microsoft, Google, and Meta that operate massive, globally distributed data centre infrastructures at unprecedented scale to deliver cloud computing services.
- **Inference**: The process of using a trained machine learning model to make predictions or generate outputs on new, unseen data, representing the operational phase after model training is complete.
- **Mixture of Experts**: A machine learning architecture that uses multiple specialized sub-models (experts) and a gating network to determine which experts should process each input, improving efficiency and performance for large models.
- **MLOps**: Machine Learning Operations - a set of practices that automate and streamline the machine learning lifecycle from development to deployment, monitoring, and maintenance, combining ML, DevOps, and data engineering principles.

- **Model Collapse**: A phenomenon where generative AI models degrade over time when trained on their own outputs or synthetic data, leading to reduced diversity and quality in generated content.
- **Multimodal AI**: AI systems capable of processing and understanding multiple types of data inputs simultaneously, such as text, images, audio, and video, to perform complex reasoning and generation tasks.
- **Noisy Neighbours**: A cloud computing issue where one tenant's resource-intensive workloads negatively impact the performance of other tenants sharing the same physical infrastructure, affecting application performance and user experience.
- **Overfitting**: An undesirable machine learning behaviour where a model learns the training data too closely, including noise and irrelevant details, resulting in poor performance on new, unseen data.
- **Pass-through clause**: A contractual provision that requires one party to extend certain obligations or requirements from its own contract to another party, typically used in subcontracting to ensure compliance with original terms.
- **PEFT**: Parameter-Efficient Fine-Tuning - techniques that adapt large pre-trained models to new tasks by training only a small subset of parameters while keeping most of the original model frozen, reducing computational costs and training time.
- **Perplexity signal**: A measure of predictability in text that helps identify AI-generated content, where lower perplexity (more predictable) text is more likely to be AI-generated, while higher perplexity suggests human authorship.
- **Power density**: The amount of electrical power consumed per unit of space in data centres, typically measured in kilowatts per rack, with modern AI workloads driving requirements to 50-100+ kW per rack.
- **Prompt Injection**: A cybersecurity exploit where adversaries craft inputs to manipulate AI language models into behaving in unintended ways, bypassing safety measures and potentially causing unauthorized actions.
- **Prompt Leakage**: A security vulnerability where AI models inadvertently reveal their system prompts or sensitive instructions through their outputs, potentially exposing proprietary information or attack vectors.
- **Retrieval-Augmented Generation**: A technique that enhances large language models by combining them with external knowledge retrieval systems, allowing models to access up-to-date information and provide more accurate, grounded responses.
- **Synthetic Data**: Artificially generated data that mimics the statistical properties and patterns of real-world data, used for training AI models, testing systems, and protecting privacy while maintaining data utility.

- **traffic loss rates**: The percentage of data packets that are lost or dropped during network transmission due to congestion, hardware failures, or capacity limitations, impacting network performance and application quality.
- **Transformative agreements**: Contracts between institutions and publishers that transform the business model of scholarly publishing from subscription-based access to open access publishing, typically combining reading access with publishing fees.
- **Underfitting**: A machine learning scenario where a model is too simple to capture the underlying patterns in data, resulting in poor performance on both training and test data due to high bias and oversimplified assumptions.
- **Vacancy rate**: The percentage of available data centre capacity that is unoccupied, typically measured in megawatts or rack space, used as a key market indicator for supply and demand dynamics.
- **Vector Database**: A specialized database designed to store, index, and search high-dimensional vector embeddings efficiently, enabling similarity search and powering AI applications like semantic search and recommendation systems.



Peter G. Osborn
pgo@peterosborn.com
+44 (0)7802-666758
https://www.peterosborn.com