AI Infrastructure: Key Terminologies Cheat Sheet

■■ Compute (Processing Power)

- GPU (Graphics Processing Unit) → Specialized chip for parallel processing, critical for AI.
- TPU (Tensor Processing Unit) → Google’s custom chip for ML workloads.
- Accelerators → Specialized hardware (ASICs, FPGAs) for speeding up AI tasks.
- Cluster → Group of connected machines working together.
- Node → A single machine in a cluster.
- Distributed Training → Splitting model/data across multiple GPUs/nodes.

■ Data & Storage

- Dataset → The collection of data (text, images, logs) used for training/testing.
- Data Lake → Centralized raw data storage system.
- Data Pipeline → Automated flow to collect, clean, transform, and feed data into models.
- Sharding → Splitting large data into smaller chunks for parallel processing.

■ Networking

- Bandwidth → How much data can flow per second (Gbps).
- Latency → Delay in transferring data (lower is better).
- InfiniBand / NVLink → High-speed interconnects for GPU clusters.

■■ Virtualization & Orchestration

- Virtual Machine (VM) → Simulated computer environment.
- Container (Docker) → Lightweight, portable environment for apps/models.
- Kubernetes (K8s) → Orchestrates (manages, scales) containers across clusters.
- Scheduler → Assigns jobs to resources (GPUs, CPUs).

■■ Cloud & Scaling

- On-premises → Hardware hosted locally (your own data center).
- Cloud (AWS, Azure, GCP) → Renting infra on demand.
- Hybrid Cloud → Mix of on-prem + cloud.
- Elasticity → Ability to scale resources up or down instantly.
- Serverless → Running workloads without managing servers.

■ Model Training & Deployment

- Training → Teaching a model from data.
- Inference → Using a trained model to make predictions.
- Serving → Deploying models to respond to real-world queries.
- Load Balancer → Distributes requests across servers.
- Latency Optimization → Making inference as fast as possible.

■ Observability & Optimization

- Monitoring → Tracking performance of infra & models.
- Logging → Recording system and application events.
- Profiling → Checking how resources (GPU, memory) are used.
- Autoscaling → Automatic adjustment of resources.
- Cost Optimization → Reducing GPU/cloud usage cost.