AI Fundamentals CHEAT SHEET

Plex Consulting

Types of AI

Type	Description
Narrow AI	Designed for a single task. All current AI is narrow — e.g. image recognition, translation, chatbots.
General AI (AGI)	Hypothetical AI that can perform any intellectual task a human can. Does not exist yet.
Super AI (ASI)	Theoretical AI surpassing human intelligence in all domains. Purely speculative at this stage.

Open vs Closed AI

Model	Description
Open Source	Weights and code publicly available. Run locally, fine-tune, customise. No data leaves your environment. E.g. LLaMA, Mistral, Phi, Qwen.
Closed Source	Proprietary models via API or web interface. Provider controls infrastructure and data handling. E.g. GPT-4, Claude, Gemini.

On-Premises AI Options

Option	Description
On-Prem RAG	Documents indexed locally; an LLM searches them to answer questions. Data stays internal.
Full On-Prem LLM	Open-source model on your own hardware. Maximum control and privacy, requires GPU infrastructure.
Cloud API	Prompts sent to a cloud provider. Simplest to deploy, but data transits external servers.

Generative vs Non-Generative AI

Type	Description
Generative AI	Creates new content — text, images, code, audio, video. E.g. ChatGPT, Claude, Midjourney, Copilot.
Non-Generative	Classifies, predicts, or optimises based on existing data. E.g. spam filters, fraud detection, recommendations.

Techniques That Power AI

Technique	What It Does
Machine Learning	Learns patterns from data to make predictions without explicit programming.
Deep Learning	Neural networks with many layers handling complex, unstructured data.
NLP	Enables machines to understand, interpret, and generate human language.
Reinforcement Learning	AI learns by trial and error, receiving rewards for correct actions.
Computer Vision	Machines interpret visual information from images and video.

How Large Language Models Work

Training Phase

1. **Data collection** — billions of text documents from the internet, books, code, and curated datasets.
2. **Tokenisation** — text broken into tokens (words or sub-words). Vocabulary of 30,000—100,000 tokens.
3. **Pattern learning** — the model learns statistical relationships between tokens by predicting the next word.
4. **Fine-tuning** — refined on curated examples and human feedback (RLHF) to improve helpfulness and safety.
5. **Evaluation** — benchmarked for accuracy, reasoning, safety, and bias before release.

Using an LLM (Inference)

1. **You send a prompt** — your text is tokenised and sent to the model along with system instructions.
2. **Context window applies** — the model can only "see" a fixed number of tokens at once (e.g. 128K for GPT-4).
3. **Token-by-token generation** — the model predicts one token at a time, each informed by preceding tokens.
4. **Temperature controls randomness** — low = more predictable; high = more creative and varied output.
5. **Output returned** — generated tokens decoded back into human-readable text.