Understanding Artificial Intelligence & Large Language Models

1. What is Artificial Intelligence (AI)?

Artificial Intelligence is the simulation of human intelligence in machines that are programmed to think, reason, and make decisions. AI spans various domains including computer vision, natural language processing (NLP), robotics, and expert systems.

2. Branches of AI

- Machine Learning (ML)
- Deep Learning (DL)
- Natural Language Processing (NLP)
- Computer Vision
- Expert Systems
- Robotics

3. Machine Learning (ML)

Machine Learning is a subset of AI focused on algorithms that can learn from and make predictions on data.

- Supervised Learning: Labeled data, e.g., classification, regression.
- Unsupervised Learning: No labels, e.g., clustering.
- Reinforcement Learning: Agents learn by rewards/punishments.

4. Deep Learning and Neural Networks

Deep Learning uses neural networks with multiple layers (hence "deep").

- ANN (Artificial Neural Network)
- CNN (Convolutional Neural Network) - for images.

- RNN (Recurrent Neural Network) - for sequences.

5. Natural Language Processing (NLP)

NLP allows machines to understand, interpret, and generate human language.

- Tasks: Text classification, Named Entity Recognition, Machine Translation, Summarization.
- Preprocessing: Tokenization, Stopword removal, Lemmatization.

6. Transformers - The Foundation of Modern LLMs

Transformers are deep learning models introduced in the paper "Attention is All You Need" (Vaswani et al., 2017).

Key Concepts:

- Self-Attention: Allows the model to weigh the importance of different words in a sequence.
- Positional Encoding: Since Transformers don't have recurrence, they need position info.
- Encoder & Decoder: Used in sequence-to-sequence tasks.

7. How a Transformer is Built (from Scratch)

- Input Embeddings: Map words to vectors.
- Add Positional Encodings.
- Pass through multiple encoder blocks:
- Each block has Multi-Head Attention, Add & Norm, Feed-Forward, Add & Norm.
- Decoder blocks are similar but include masked attention and cross-attention.
- Final output is passed through a linear layer and softmax.

8. Training Large Language Models (LLMs)

- Datasets: Massive corpora (Common Crawl, Wikipedia, Books).
- Objective: Predict next token (causal LM) or masked token (masked LM).
- Optimizers: Adam, AdamW.

- Loss Function: Cross-Entropy Loss.

- Hardware: GPUs/TPUs, distributed training.

9. Popular LLM Architectures

- GPT (Generative Pre-trained Transformer): Decoder-only model.

- BERT (Bidirectional Encoder Representations from Transformers): Encoder-only.

- T5 (Text-to-Text Transfer Transformer): Encoder-decoder.

- LLaMA, Falcon, PaLM, etc.

10. Applications of AI and LLMs

- Chatbots (e.g., ChatGPT)

- Code generation

- Image captioning

- Search engines

- Personal assistants

- Content recommendation

11. Challenges in AI

- Bias and fairness

- Interpretability

- Data privacy

- High computation cost

12. Future of AI

- Better multi-modal models

- On-device AI

- Improved interpretability and ethics

- Integration with robotics and IoT

This document provides a high-level overview of foundational AI concepts, especially LLMs and Transformers. For practical implementations, one should explore open-source projects like Hugging Face Transformers, LangChain, and DeepSpeed.