# Hands-On Large Language Models - Notes (Ch. 1-3)

## Chapter 1: Introduction to Large Language Models

LLMs revolutionized Language AI by transforming translation, summarization, and text generation tasks.
ChatGPT, built on GPT-3.5, marked the global boom of conversational AI.

HISTORY
- Bag-of-Words: simple frequency representation.
- Word2Vec (2013): learned word meanings as vectors.
- Attention (2014): focus on relevant tokens.
- Transformer (2017): introduced parallel self-attention.

MODELS
- BERT (Encoder-only): understands text.
- GPT (Decoder-only): generates text.
- T5 (Encoder-Decoder): translation and summarization.

TRAINING PARADIGM
1. Pretraining on huge corpora.
2. Fine-tuning for specific tasks.
3. RLHF for alignment with human feedback.

ETHICS
Bias, misinformation, and data privacy remain critical issues.

## Chapter 2: Tokens and Embeddings

Tokens and embeddings are core to how LLMs understand language.

TOKENIZATION

- Word-level, Subword-level (BPE, WordPiece), Character-level.
- Adds [CLS], [SEP], <BOS>, <EOS> for structure.

EMBEDDINGS

- Word2Vec: static, semantic vectors.
- Contextual embeddings (BERT, GPT): dynamic meanings.
- Sentence embeddings: represent full text for search or clustering.

Visualization (via PCA/UMAP) shows similar words cluster close together.

## Chapter 3: Inside Large Language Models

Transformers are the foundation of all modern LLMs.

ARCHITECTURE

Each block has:

1. Multi-head self-attention
2. Feed-forward layer
3. Normalization & residuals

ATTENTION

Attention(Q,K,V) = softmax(QK^T / sqrt(d_k)) * V

ENCODERS vs DECODERS

Encoders (BERT) read all tokens; Decoders (GPT) predict next tokens one by one.

TRAINING

- Loss: Cross-entropy.
- Optimizer: AdamW.
- Data: large-scale web corpora.

EFFICIENCY

- FlashAttention, Quantization, LoRA, Mixture of Experts.

SCALING

Performance scales with parameters, data, and compute.