Session 2.

Demystifying LLMs – from ‘black boxes’ to ‘white boxes’

Understanding the Architecture of Large Language Models

Prof. Ken Spours, College of Education, Capital Normal University
& UCL Institute of Education

Aims of Session 2

- Understand recent evolution of LLMs and their technological trajectory
- Understand how LLMs work internally - their technological architecture
- Recognise the differences between Chinese (e.g. DeepSeek) and US models
- Relate the internal architecture to creative working with the machine

Collaborative Critical Praxis

- LLM output is not magic - its quality and safety are directed by its mechanics and control layers – its **Technological Architecture**
- Understanding the deep LLM architecture and pre-training so *users* can exercise greater control through **Collaborative Critical Praxis (CCP)**
- **CCP is combinational/organic** - increasing **breadth** of general political-economy-ecology understanding **+** increasing **depth** of specialist knowledge of LLMs
- CCP drives demystification - turning LLM **black boxes** (opaque) into **white boxes** (transparent).

Recent evolution of general & specialist LLMs

Evolution step	Technical focus	Purpose & control
Scaling (e.g. GPT-4, Qwen 3 Max)	Increasing the number of parameters and the size of the training data (billions of tokens).	Unlocking 'emergent abilities' (e.g., coding, translation).
Alignment and machine learning	Instruction Fine-Tuning and Reinforcement- Learning from Human Feedback.	Teaching the raw, pre-trained model to follow explicit user instructions - the source of the model's 'personality' and ethical guardrails.
Multi-modality (e.g., Co-pilot, Gemini, GPT-4/5, Qwen-VL+)	Training models on diverse data streams: text, images, audio, and video.	Enabling the model to understand and generate content across different data types
Specialist LLMs (e.g. Engineering, Medicine)	Domain-specific LLMs.	Deeper contextual understanding, enhancing reliability for professional use.

What are LLMs and how do they work?

Definition - an LLM is a large, pre-trained sequence-to-sequence GenAI model.

Function - fundamentally a predictive text engine, calculating the probability of the next word (comprised of tokens of text).

Scale - ‘large’ based number of parameters & dataset size.

Token probability - all LLM outputs, even complex ones, are based on token probability.

Token – the fundamental unit

A token is a fragmented unit of text—it can be a full word, part of a word (e.g., "ing"), punctuation, or even a single character and a preceding space.

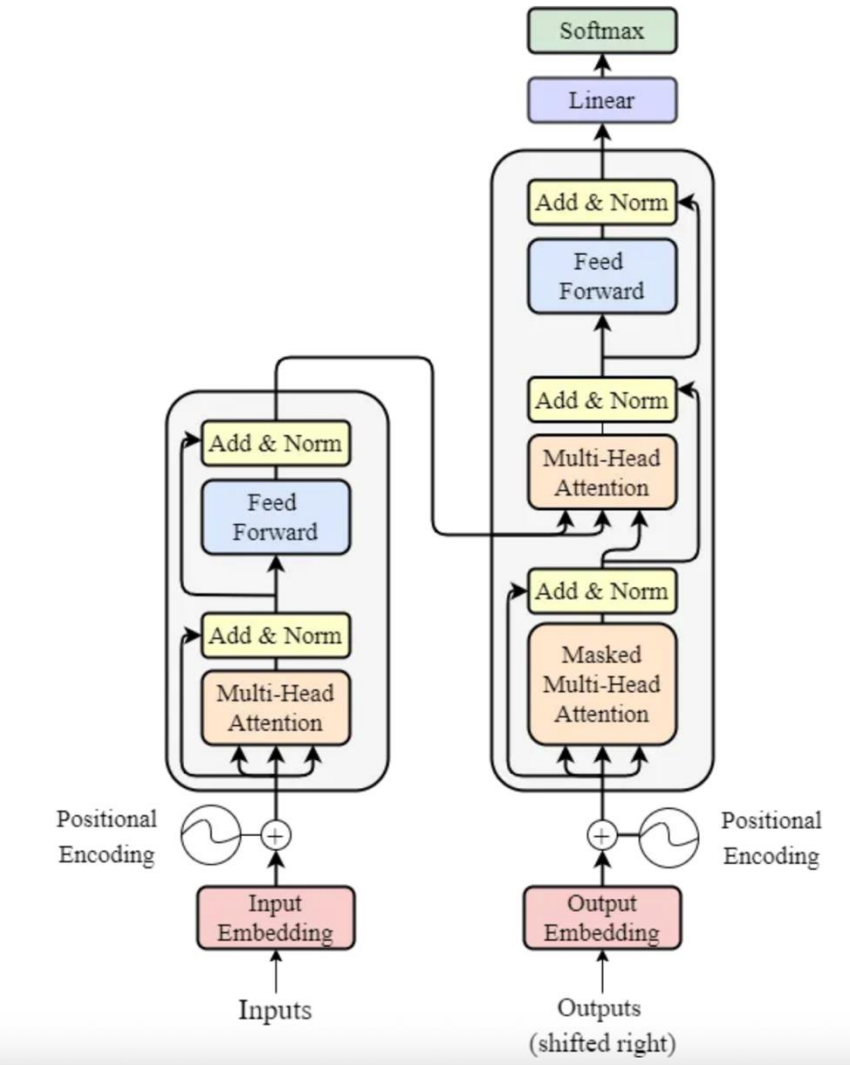
Tokenization process - all input text (prompts) and output text (responses) are first broken down by a **tokenizer** into these fixed, numerical tokens.

Numerical representation (embedding) - each unique token is permanently mapped to a specific numerical vector (a list of numbers) called an **embedding** that the LLM's mathematical circuits can understand.

Model size and cost - the size of an LLM's **vocabulary** is the total number of unique tokens it recognizes (often 50,000 - 100,000). The model's computational cost (and speed) is directly proportional to the number of tokens it processes.

Context window limit - the **context window** (or maximum input length) is measured in tokens (e.g., 8,000, 128,000, 1M) - limiting how much past conversation the model can **remember** and process at any one time.

Multi-Head Attention Unit

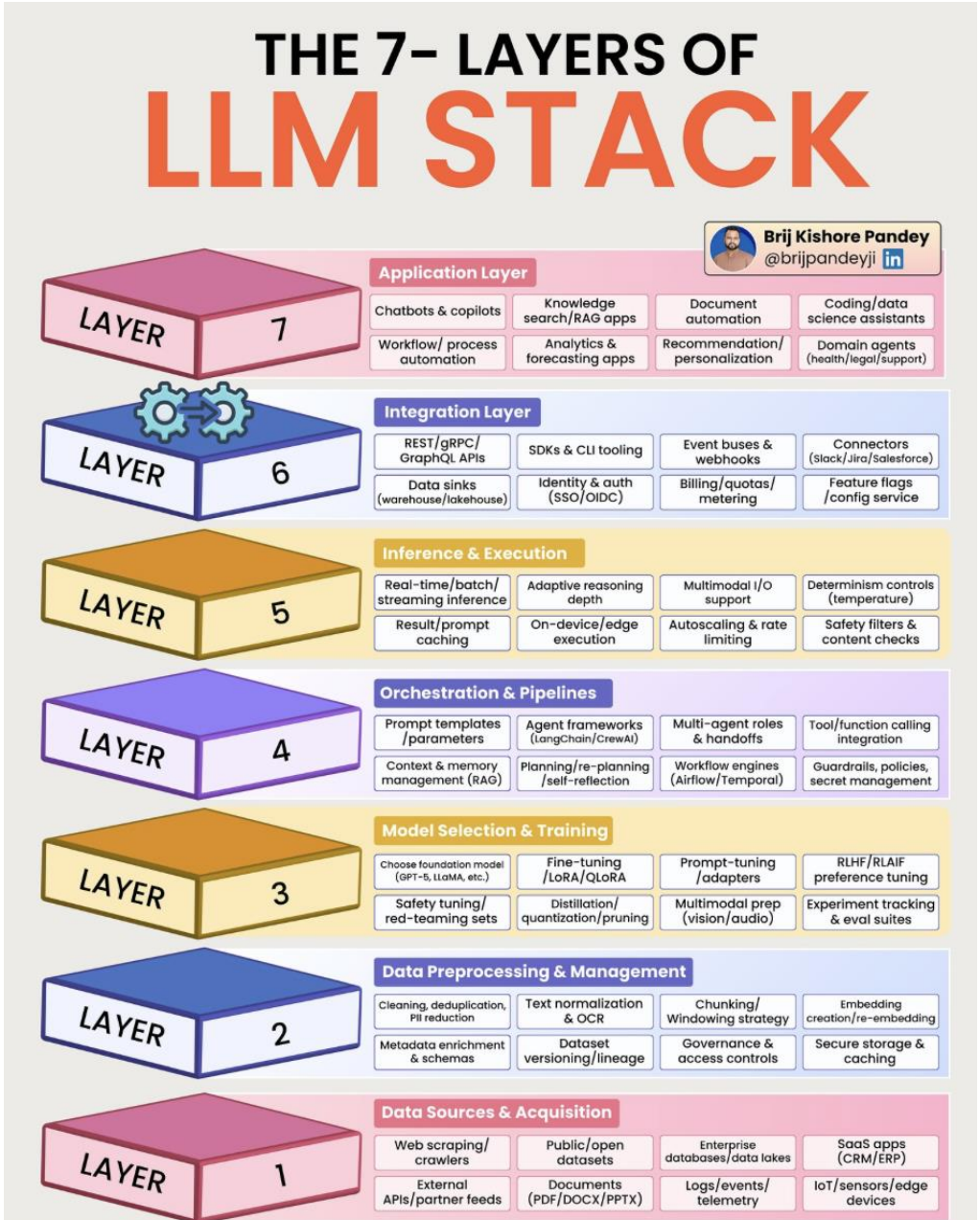


Key innovation – parallelism - all data tokens *simultaneously* instead of word-by-word.

Stacked technological structure - each stack contains a ***Multi-Head Attention Unit*** and a simple Feed-Forward Network – sub-component of **transformer blocks.**

Central mechanism - Self-Attention Mechanism to calculate relationships and context across long sequences.

What makes LLMs large



LLMs are **deep networks** of **transformer blocks** (or Layer) – often over 100 layers.

The Multiplier - Total **Parameters** (the model's size, e.g., 100B) = Depth (Layers, L) and **Width** (Hidden Size, H) - the data width used.

The Cost - increasing the **Width** (H) results in a non-linear cost increase.

Layer 5 – is the component that produces **depth**

The Challenge - architectures like DeepSeek-V3's **Mixture-of-Experts** (MoE) show that high performance can be achieved by **sparsifying** the block, challenging the brute-force scaling of dense Western models.

Expert Feedforward Network (FFN) – DeepSeek employs ‘expert’ FFNs that use only 2-5 per cent of the transformer to produce outputs thus radically reducing cost.

Self-attention mechanism

Concept	Explanation	Practical Outcome
Dynamic weighing	As the model processes a single word in a sentence (e.g., the word "it"), it simultaneously calculates a "score" or weight for every other word in that sentence (e.g., "The cat sat on the mat and *it* purred.").	The model does not treat all words equally, focusing only on the most relevant contextual clues.
Establishing context	The calculation determines which surrounding words are most relevant to the meaning of the current word.	Linking words over long distances creates deep contextual understanding, preventing errors and ambiguity
Parallel processing	Unlike sequential RNNs, which had to Transformer uses Attention to process the entire input sequence **at the same time**.	This parallel computation drastically speeds up training and inference, allowing the creation of models with billions of parameters (the "Large" in LLM).

Instruction Tuning (SFT) and Alignment (RLHF)

How LLMs learn their 'personality', follow instructions and apply safety guardrails.

Base LLM - predicts the next statistically likely word (often hallucinating).

Training - need to train the model to be a helpful **assistant** that follows instructions and safety rules.

The training process - Reinforcement Learning from Human Feedback (RLHF)

- ***Instruction tuning* -** teaches the model *how* to follow commands using curated examples – this takes place away from the user.

- ***Reward model* -** separate model to learn human preferences - reviewers rating outputs (best to worst).

- ***Final tuning*** - the main LLM is tuned using this Reward Model, rewarding outputs that are highly human-aligned.

System prompt – the superior command

The **‘hidden prompt’** and control parameters as immediate, practical levers for output style and reliability.

- **Definition -** hidden instruction (or "prime directive") given to the LLM *before* the user's first query.
- **Purpose -** stablishes the model's **persona, role, rules, and constraints** for the entire conversation.
- **Control Mechanism - o**verrides general training to enforce specific behaviors, such as tone ("Act as a friendly tutor") or format ("Respond only in JSON").
- **Superior to User Prompts - h**olds more weight than standard user input because the LLM is explicitly trained (via fine-tuning/RLHF) to prioritize the System Prompt's instructions.
- **Practical use - e**ssential for building reliable, consistent applications on top of base LLMs, turning them from general chatbots into specialized tools.

The LLM Reward Model – problem or solution?

- **Proprietary Data Collection - t**he Reward Model is trained on the highly subjective and sensitive data of **human preferences** (RLHF data). The RLHF imprints the values of the LLM.

- **Rating initial outputs -** the tech company pays human contractors (often offshore) to rate thousands of model outputs for helpfulness and harmlessness. This dataset of preferences is a critical, competitive asset and is not publicly shared.

- **Internal feedback loop -** the Reward Model never leaves the development environment - its sole function is to continuously evaluate and guide the larger LLM during its fine-tuning phase.

- **Control and policy -** the Reward Model determines what the LLM defines as ‘good’ or ‘bad’ behaviour - it is the primary mechanism for implementing the company's safety policies and ideological orientation.

- The role of user **critical control practice** in relation to the LLM Reward Model

DeepSeek-V3 vs. US Models: Strategic Differences

D-V3 (China) Focus	US Models (Western) Focus	Critical Significance (CTP)
Architectural Efficiency (Low active cost, MoE innovations like MLA)	**Brute-Force Scaling** (High parameter count, massive compute budgets)	**Innovation Under Constraint:** DeepSeek's design is a material response to hardware sanctions, achieving SOTA performance at a fraction of the cost.
Open Weights (MIT License, fully commercial)	**Closed/Restricted API Access** (Proprietary control)	**Democratization:** Challenges US dominance by lowering the barrier to entry for developers globally.
Bilingual/Local Data Bias (Higher Chinese proportion)	**English/Western Data Bias** (Global web/code focus)	**Cultural/Geopolitical Alignment:** Ensures superior performance and cultural fit for non-Western linguistic domains.
State Regulatory Alignment (Mandatory ideological/policy guardrails)	**Corporate Policy Alignment** (Company safety/ethical guidelines)	**Ideological Control:** The underlying **Reward Model** implements fundamentally different political and economic controls on output.

End of Session Quiz (1)

1. The core function of an LLM

a. An advanced semantic search engine.
b. A context-aware predictive text engine that calculates the probability of the next token.
c. A deterministic algorithm for generating factual knowledge.
d. A word-frequency counter.

2. The unit of text processing

a. A parameter.
b. An embedding.
c. A token.
d. A layer.

3. The 'Large' in LLM

a. The depth of the attention mechanism and the width of the training data.
b. The number of layers and the size of the context window.
c. The number of parameters and the size of the training dataset (in tokens).
d. The speed of inference and the number of expert networks.

End of Session Quiz (2)

4. Key Innovation of the Transformer Architecture

a. The use of a Recurrent Neural Network (RNN) structure.
b. Parallel processing of the input sequence.
c. Sequential, word-by-word processing.
d. The automatic translation of text into images.

5. The Purpose of Alignment (e.g. Instruction Fine-Tuning and RLHF)

a. Drastically increase the number of model parameters.
b. Train the model to follow explicit instructions.
c. Enable the model to process images and audio data.
d. Unlock the model's emergent abilities.

6. The Role of the Reward Model

a. Serves as a public, open-source record of all human-rated preferences.
b. Directly executes the user's prompt during inference.
c. Implements the developer's safety policies and ideological orientation.
d. Tracks the model's performance in language translation tasks.

End of Session Quiz (3)

8. The System Prompt

a. The user types it first.
b. The LLM is explicitly trained (via fine-tuning/RLHF) to prioritize its instructions,
c. It is an optional feature.
d. It increases the context window limit.

9. Mixture-of-Experts (MoE) model is responsible for deciding which specialized sub-networks will process a specific input token?

a. The Multi-Head Attention Unit.
b. The Softmax Layer.
c. The Router (or Gating Network).
d. The Positional Encoding Layer.

Preparation for Session 3

Preparation for Session 3 - please read the following texts.

Atchley, P., et al. (2024) Human and AI collaboration in the higher education environment: opportunities and concerns *Cognitive Research: Principles and Implications* https://doi.org/10.1186/s41235-024-00547-9

Lee, H.-P., et al., (2025) 'The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers.' In: *CHI '25: Proceedings of the CHI Conference on Human Factors in Computing Systems,* Yokohama, Japan, 26 April–1 May 2025. ACM, pp. 1–23. Available at: https://doi.org/10.1145/3706598.3713778.