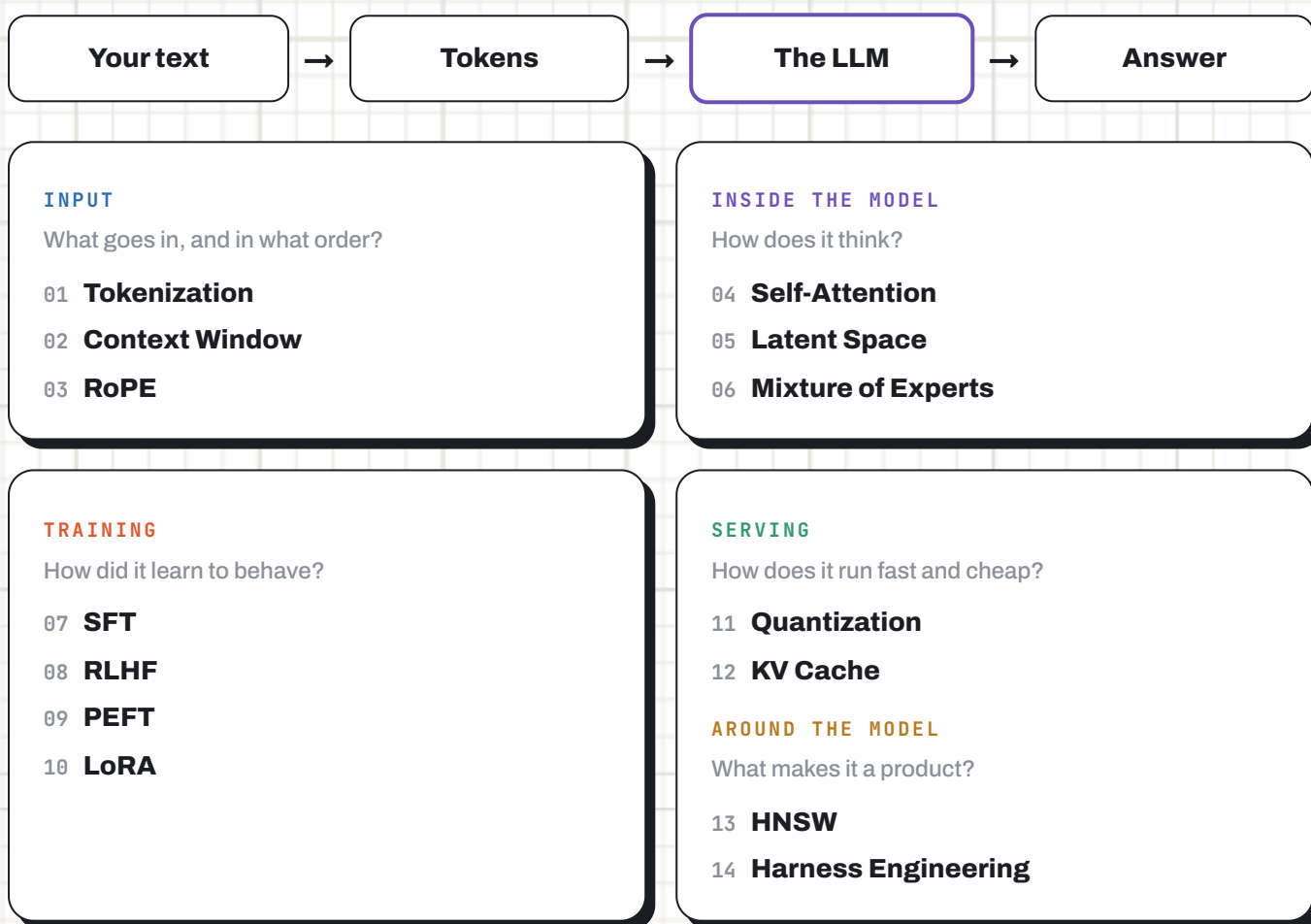


14 LLM Concepts

EVERY AI ENGINEER SHOULD KNOW

What each one is, why it matters in a real system, and how to say it in one line in an interview.



SUMIT KUMAR DASH
SKDNEURON.COM • @SKDNEURON

SAVE FOR YOUR NEXT INTERVIEW

The cheat sheet.

ONE LINE EACH

#	CONCEPT	WHERE	IN ONE LINE
01	Tokenization	INPUT	Text is split into tokens; the model only ever sees token IDs.
02	Context Window	INPUT	The most tokens the model can take in at once.
03	RoPE	INPUT	Encodes each token's position by rotating its vectors.
04	Self-Attention	MODEL	Every token weighs every other token to pick up context.
05	Latent Space	MODEL	The vector space where similar meanings sit close together.
06	Mixture of Experts	MODEL	A router sends each token to a few expert sub-networks.
07	SFT	TRAINING	Teaches behaviour from example prompts and ideal answers.
08	RLHF	TRAINING	Aligns the model with human preferences via a reward model.
09	PEFT	TRAINING	Fine-tunes a small set of parameters, freezes the rest.
10	LoRA	TRAINING	The popular PEFT method: a small low-rank update on frozen weights.
11	Quantization	SERVING	Stores weights in fewer bits to cut memory.
12	KV Cache	SERVING	Saves past keys and values so each new token is cheap.
13	HNSW	SYSTEM	A layered graph index for fast approximate vector search.
14	Harness Engineering	SYSTEM	The system around the model that makes it reliable.

How to use this guide: read the cheat sheet, then the card for anything you couldn't explain out loud. Each card ends with a one-line answer you can use in an interview, and a link to the original paper.

01 Tokenization

INPUT

WHAT IT IS

Splitting text into tokens (whole words, word pieces or characters), each mapped to an ID. Most LLMs use subword schemes such as byte pair encoding (BPE).

WHY IT MATTERS

Models see tokens, not letters, which is why they stumble at spelling and counting characters. Prices and limits are counted in tokens, and many languages need more tokens per word than English.

SAY IT **"The model never sees text, only token IDs. Cost, limits and some odd failures all come from that."**

Sennrich et al., 2015 · Neural Machine Translation of Rare Words with Subword Units · arxiv.org/abs/1508.07909

02 Context Window

INPUT

WHAT IT IS

The maximum number of tokens the model can take in at once, usually counting both your input and its output.

WHY IT MATTERS

Anything beyond it is cut off. Longer inputs cost more and run slower, and models use the middle of long inputs worse than the start or end.

Best **"at the beginning or end of the input context"**

SAY IT **"A bigger window isn't free: more cost, more latency, and the middle gets ignored. Retrieve less, but better."**

Liu et al., 2023 · Lost in the Middle: How Language Models Use Long Contexts · arxiv.org/abs/2307.03172

03 RoPE Rotary Position Embedding

INPUT

WHAT IT IS

Tells the model where each token sits by rotating its query and key vectors by an angle that depends on position, so attention reflects how far apart two tokens are.

WHY IT MATTERS

Attention alone ignores word order. RoPE is used in many open models, such as the LLaMA family, and many long-context tricks work by rescaling it.

SAY IT **"RoPE turns position into a rotation, so attention depends on the distance between tokens."**

Su et al., 2021 · RoFormer: Enhanced Transformer with Rotary Position Embedding · arxiv.org/abs/2104.09864

04 Self-Attention

MODEL

WHAT IT IS

Each token builds a query, a key and a value. It scores every other token (query $\times$ key) and takes a weighted mix of their values: $\text{softmax}(QK^T/\sqrt{d}) \cdot V$.

WHY IT MATTERS

This is how context flows between words, so "it" can find "the bot". Compute grows with the square of the sequence length, which is why long contexts get expensive.

SAY IT "Every token looks at every other token and decides what to borrow. That's why cost is quadratic in length."

Vaswani et al., 2017 · Attention Is All You Need · arxiv.org/abs/1706.03762

05 Latent Space

MODEL

WHAT IT IS

The model's internal vector space, where text is represented as points (embeddings). Things with similar meaning end up close together.

WHY IT MATTERS

It's what makes semantic search work: "refund policy" lands near "money-back rules" with no words in common. Embedding models and vector databases are built on it.

SAY IT "Meaning becomes geometry. Similar ideas are nearby vectors, so search becomes a nearest-neighbour lookup."

Mikolov et al., 2013 · Efficient Estimation of Word Representations in Vector Space (word2vec) · arxiv.org/abs/1301.3781

06 MoE Mixture of Experts

MODEL

WHAT IT IS

Replaces one big feed-forward block with many smaller "experts", plus a router that sends each token to only a few of them.

WHY IT MATTERS

The knowledge of a huge model at the per-token compute of a much smaller one. Memory still has to hold every expert.

Mixtral: 47B parameters available, 13B active per token

SAY IT "Large total parameters, small active parameters: a router picks a few experts for each token."

Shazeer et al., 2017 · The Sparsely-Gated Mixture-of-Experts Layer · arxiv.org/abs/1701.06538
Example: Jiang et al., 2024 · Mixtral of Experts · arxiv.org/abs/2401.04088

07 SFT Supervised Fine-Tuning

TRAINING

WHAT IT IS

Fine-tuning a pre-trained model on example prompts paired with ideal answers.

WHY IT MATTERS

It turns a text-completer into an assistant that follows instructions and formats. Usually the first step after pre-training.

SAY IT "SFT teaches behaviour by example: show the model the answers you want."

Wei et al., 2021 · Finetuned Language Models Are Zero-Shot Learners (FLAN) · arxiv.org/abs/2109.01652

08 RLHF Reinforcement Learning from Human Feedback

TRAINING

WHAT IT IS

People rank answers, a reward model learns their preferences, and the LLM is optimised with RL to score well.

WHY IT MATTERS

It's how assistants learn qualities that are easier to judge than to write examples for: helpfulness, tone, safety.

SAY IT "SFT shows the right answer. RLHF teaches which of two answers is better."

Ouyang et al., 2022 · Training language models to follow instructions with human feedback · arxiv.org/abs/2203.02155

09 PEFT Parameter-Efficient Fine-Tuning

TRAINING

WHAT IT IS

Fine-tuning that freezes most of the model and trains only a small set of added or selected parameters.

WHY IT MATTERS

Fine-tuning fits on modest GPUs, and each task becomes a small file instead of a full copy of the model.

SAY IT "Train a tiny slice, freeze the rest: most of the gain at a fraction of the cost."

Lialin et al., 2023 · Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning · arxiv.org/abs/2303.15647

10 LoRA Low-Rank Adaptation

TRAINING

WHAT IT IS

The most popular PEFT method: adds a small low-rank update beside the frozen weights and trains only that.

WHY IT MATTERS

Cheap to train, and the update can be merged back in, so serving costs nothing extra.

GPT-3 175B: 10,000× fewer trainable parameters

SAY IT "LoRA learns a small low-rank 'delta' on top of frozen weights."

Hu et al., 2021 · LoRA: Low-Rank Adaptation of Large Language Models · arxiv.org/abs/2106.09685

11 Quantization

SERVING

WHAT IT IS

Storing weights in fewer bits, for example 16-bit down to 8-bit or 4-bit.

WHY IT MATTERS

Memory drops roughly in proportion: a 7B model's weights need about 14 GB at 16-bit and about 3.5 GB at 4-bit, usually for a small quality cost.

LLM.int8(): half the memory, full-precision results

SAY IT **"Fewer bits per weight, so the model fits on smaller GPUs, trading a little accuracy for memory."**

Dettmers et al., 2022 · LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale · arxiv.org/abs/2208.07339

12 KV Cache Key-Value Cache

SERVING

WHAT IT IS

While generating, the model saves the keys and values it computed for earlier tokens, so each new token only computes its own.

WHY IT MATTERS

Without it, every step would redo the whole text. It grows with context length and users, so long chats get slow and costly.

PagedAttention (vLLM): 2–4× higher throughput

SAY IT **"The KV cache stops the model recomputing the past. Its memory is often the serving bottleneck."**

Kwon et al., 2023 · Efficient Memory Management for LLM Serving with PagedAttention · arxiv.org/abs/2309.06180

13 HNSW Hierarchical Navigable Small World

SYSTEM

WHAT IT IS

A graph index for approximate nearest-neighbour search. Vectors link to their neighbours in several layers; search starts in the sparse top layer and zooms in.

WHY IT MATTERS

It's how most vector databases search millions of embeddings quickly for RAG, trading a little recall for a lot of speed. Search-breadth settings control that trade.

SAY IT **"Layers of neighbour graphs, searched coarse to fine: approximate, but very fast."**

Malkov & Yashunin, 2016 · Efficient and robust approximate nearest neighbor search... · arxiv.org/abs/1603.09320

14 Harness Engineering

SYSTEM

WHAT IT IS

Everything built around the model to make it dependable: prompts, tool calls, retries, output validation, guardrails, memory and evaluations.

WHY IT MATTERS

The model is one unreliable component; the harness is what makes the product reliable. A better harness often beats a bigger model.

SAY IT **"Reliability comes from the system around the model, not from the model alone."**

Watch: SKD Neuron · Harness Engineering, the full video · youtube.com/watch?v=WRm9T10C70s

One question, all 14 concepts.

WHAT HAPPENS WHEN YOU ASK

- 01 You ask a support bot: "What's your refund window?" The **harness** 14 takes over: it builds the prompt, decides to search your documents, and will check the answer before you see it.
- 02 Your question is turned into an embedding, a point in **latent space** 05. **HNSW** 13 finds the nearest document chunks in a fraction of a second.
- 03 The prompt and chunks are **tokenized** 01, and all of it has to fit in the **context window** 02. Put the best chunk first or last, not in the middle.
- 04 In every layer, **self-attention** 04 lets each token pull context from the others, **RoPE** 03 tells it where each token sits, and in **MoE** 06 models a router sends each token to a few experts.
- 05 The answer is generated one token at a time. The **KV cache** 12 keeps earlier work so each new token is cheap, and **quantization** 11 is what let the model fit on the GPU in the first place.
- 06 It answers like a helpful assistant because of **SFT** 07 and **RLHF** 08, and it knows your company's tone because of a **LoRA** 10 adapter, one kind of **PEFT** 09.

If the answer is wrong, these 14 also tell you where to look: **retrieval** (05, 13), **input** (01, 02), the **model** (04, 06), **training** (07–10) or the **harness** (14). Architecture first.

Found this useful?

Follow SKD Neuron for more explainers on RAG, AI agents and AI system design. Save this for your next interview.

@SKDNEURON
SKDNEURON.COM