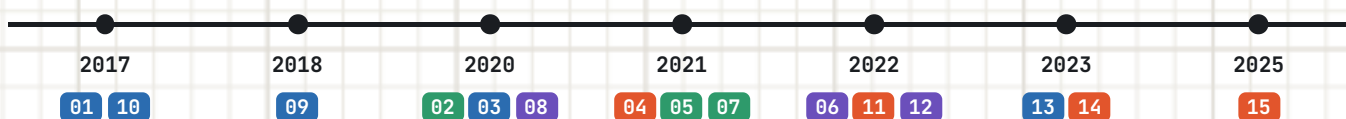


15 AI Papers

MUST READ FOR AI ENGINEERS

What each paper introduced, why it still matters, and the order to read them in. Every title links to the original.



FOUNDATIONS

How LLMs are built.

- 01 **Attention Is All You Need**
- 03 **GPT-3**
- 09 **BERT**
- 10 **Mixture of Experts**
- 13 **LLaMA**

TRAINING + ALIGNMENT

How they learn to behave and reason.

- 04 **LoRA**
- 11 **RLHF**
- 14 **DPO**
- 15 **DeepSeek-R1**

USING LLMs

Prompting, retrieval and tools.

- 06 **Chain-of-Thought**
- 08 **RAG**
- 12 **ReAct**

IMAGES

Generating and understanding pictures.

- 02 **DDPM**
- 05 **Stable Diffusion**
- 07 **CLIP**

The cheat sheet.

ONE LINE EACH

#	PAPER	YEAR	THEME	THE IDEA IN ONE LINE
01	Transformer	2017	FOUNDATION	Attention alone, no recurrence: the architecture behind every LLM.
02	DDPM	2020	IMAGES	Generate images by learning to remove noise, step by step.
03	GPT-3	2020	FOUNDATION	At 175B parameters, a few examples in the prompt replace fine-tuning.
04	LoRA	2021	TRAINING	Freeze the model, train a small low-rank update beside it.
05	Stable Diffusion	2021	IMAGES	Diffusion in a compressed latent space, steered by text.
06	Chain-of-Thought	2022	USING LLMS	Show reasoning steps in the prompt, and the model reasons better.
07	CLIP	2021	IMAGES	Images and text in one embedding space, learned from captions.
08	RAG	2020	USING LLMS	Retrieve documents first, then generate the answer from them.
09	BERT	2018	FOUNDATION	Read text in both directions by predicting masked words.
10	Mixture of Experts	2017	FOUNDATION	A router sends each token to a few experts: huge size, small compute.
11	RLHF	2022	TRAINING	Human rankings train a reward model; RL tunes the LLM to it.
12	ReAct	2022	USING LLMS	Think, act with a tool, observe, repeat: the agent loop.
13	LLaMA	2023	FOUNDATION	Smaller models, more data, open weights: the open-model wave.
14	DPO	2023	TRAINING	RLHF's goal with one simple loss: no reward model, no RL.
15	DeepSeek-R1	2025	TRAINING	Reward correct answers with RL, and reasoning emerges by itself.

How to use this guide: skim the cheat sheet, then read the card for any paper you couldn't explain out loud. Each card ends with a one-line interview answer and a link to the paper. The last page tells you which to read first.

01 Transformer Attention Is All You Need

FOUNDATION

WHAT IT INTRODUCED

An architecture built only on attention, with no recurrence or convolution. Every token looks at every other token, all in parallel.

WHY IT STILL MATTERS

GPT, BERT, LLaMA and every major LLM since are Transformers. Dropping recurrence made training parallel on GPUs, which is what made scale possible.

28.4 BLEU En→De · trained in 3.5 days on 8 GPUs

SAY IT "Attention replaced recurrence, so training went parallel, and that's what let models scale."

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

02 DDPM Denoising Diffusion Probabilistic Models

IMAGES

WHAT IT INTRODUCED

Add noise to an image step by step, and train a network to undo one step at a time. To generate, start from pure noise and keep denoising.

WHY IT STILL MATTERS

The recipe behind Stable Diffusion, Imagen and most image and video generators since. It overtook GANs on image quality.

CIFAR-10: FID 3.17, state of the art at the time

SAY IT "Train a model to remove a little noise; generating is doing that many times, starting from pure noise."

Ho, Jain & Abbeel, 2020 · Denoising Diffusion Probabilistic Models · arxiv.org/abs/2006.11239

03 GPT-3 Language Models are Few-Shot Learners

FOUNDATION

WHAT IT INTRODUCED

A 175B-parameter Transformer tested with no fine-tuning at all: you describe the task and give a few examples in the prompt.

WHY IT STILL MATTERS

It showed that scale unlocks in-context learning. That's why prompting works, and it started the prompt-engineering era.

175B parameters: 10× any earlier non-sparse model

SAY IT "Make the model big enough and the prompt becomes the program: no fine-tuning needed."

Brown et al., 2020 · Language Models are Few-Shot Learners · arxiv.org/abs/2005.14165

04 LoRA Low-Rank Adaptation

TRAINING

WHAT IT INTRODUCED

Freeze the pre-trained weights and train two small low-rank matrices beside each layer. Only those get updated.

WHY IT STILL MATTERS

Fine-tuning fits on far smaller hardware, each task is a small adapter file, and merging it back adds no latency.

vs GPT-3 full fine-tune: 10,000× fewer trainable params

SAY IT "Freeze the model, learn a small low-rank update: most of the gain at a fraction of the cost."

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

05 Stable Diffusion Latent Diffusion Models

IMAGES

WHAT IT INTRODUCED

Run diffusion in the compressed latent space of an autoencoder instead of on pixels, and add cross-attention so text can steer the image.

WHY IT STILL MATTERS

Pixel-space diffusion took hundreds of GPU days. Latents made it cheap enough to release openly as Stable Diffusion and run on one GPU.

Pixels → latent → denoise → decode

SAY IT "Same diffusion, in a much smaller latent space, so it's fast enough to run on a single GPU."

Rombach et al., 2021 · High-Resolution Image Synthesis with Latent Diffusion Models · arxiv.org/abs/2112.10752

06 Chain-of-Thought Prompting

USING LLMS

WHAT IT INTRODUCED

Prompt with examples that show the reasoning steps, not just the answers, and the model writes out its own steps before answering.

WHY IT STILL MATTERS

Big gains on maths and logic from prompting alone, in large enough models. Reasoning models like DeepSeek-R1 (15) now learn to do it by themselves.

540B + 8 examples: state of the art on GSM8K

SAY IT "Show the steps, get the steps: reasoning out loud helps big models."

Wei et al., 2022 · Chain-of-Thought Prompting Elicits Reasoning in Large Language Models · arxiv.org/abs/2201.11903

07 CLIP Contrastive Language-Image Pre-training

IMAGES

WHAT IT INTRODUCED

Train an image encoder and a text encoder together to predict which caption goes with which image, so both land in one shared space.

WHY IT STILL MATTERS

Classify images by just writing the labels as text, search images with words, and steer generators: Stable Diffusion v1 used CLIP's text encoder.

400M image-text pairs · ResNet-50 level, zero-shot

SAY IT "Images and text in one vector space, so you can search or classify images with plain words."

Radford et al., 2021 · Learning Transferable Visual Models From Natural Language Supervision · arxiv.org/abs/2103.00020

08 RAG Retrieval-Augmented Generation

USING LLMS

WHAT IT INTRODUCED

A generator paired with a neural retriever over a dense vector index of Wikipedia, so the model looks facts up instead of relying on its weights alone.

WHY IT STILL MATTERS

The pattern behind most company AI assistants: fresher knowledge, answers you can trace to a source, and facts you update without retraining.

Retriever + generator, fine-tuned together

SAY IT "Retrieve evidence first, then write from it: knowledge lives in the index."

Lewis et al., 2020 · Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks · arxiv.org/abs/2005.11401

09 BERT Bidirectional Encoder Representations from Transformers

FOUNDATION

WHAT IT INTRODUCED

An encoder-only Transformer pre-trained by hiding words and predicting them from both the left and right context, then fine-tuned with one extra layer.

WHY IT STILL MATTERS

Records on 11 NLP tasks, and "pre-train, then fine-tune" became the default. BERT-style encoders still power many embedding models and RAG rerankers.

Masked LM: hide 15% of tokens · GLUE 80.5%

SAY IT "GPT reads left to right to generate; BERT reads both ways to understand."

Devlin et al., 2018 · BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding · arxiv.org/abs/1810.04805

10 MoE Sparsely-Gated Mixture of Experts

FOUNDATION

WHAT IT INTRODUCED

Up to thousands of small "expert" networks in one layer, and a gating network that picks only a few of them for each input.

WHY IT STILL MATTERS

Parameters grow without compute per token growing with them. Mixtral and DeepSeek-V3 are MoE models, as are many large LLMs today.

Up to 137B params · 1000× more capacity

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

Shazeer et al., 2017 · Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer · arxiv.org/abs/1701.06538

11 RLHF InstructGPT

TRAINING

WHAT IT INTRODUCED

Three steps: fine-tune on human-written answers (SFT), train a reward model on human rankings, then optimise the LLM against it with RL (PPO).

WHY IT STILL MATTERS

The recipe that turned GPT-3 into an assistant that follows instructions, and the method ChatGPT was trained with.

1.3B InstructGPT preferred to 175B GPT-3

SAY IT "SFT shows the model good answers; RLHF teaches it which answers people prefer."

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

Where it started: Christiano et al., 2017 · Deep reinforcement learning from human preferences · arxiv.org/abs/1706.03741

12 ReAct Reasoning + Acting

USING LLMs

WHAT IT INTRODUCED

The model alternates Thought → Action (call a tool or API) → Observation (read the result), and loops until it can answer.

WHY IT STILL MATTERS

The basic loop inside most AI agents. Checking real tool results cuts the hallucination that plain chain-of-thought suffers from.

+34% ALFWorld · +10% WebShop, 1–2 examples

SAY IT "Reason, act, observe, repeat: that loop is what turns an LLM into an agent."

Yao et al., 2022 · ReAct: Synergizing Reasoning and Acting in Language Models · arxiv.org/abs/2210.03629

13 LLaMA Open and Efficient Foundation Language Models

FOUNDATION

WHAT IT INTRODUCED

A family of 7B–65B models trained on trillions of tokens of publicly available data only, with the weights released to researchers.

WHY IT STILL MATTERS

Showed that smaller models trained on more data can beat bigger ones, and started the open-weight wave: Llama 2 and 3, and thousands of community fine-tunes.

LLaMA-13B beats GPT-3 (175B) on most benchmarks

SAY IT "Train smaller models on more tokens and release the weights: that's how open LLMs took off."

Touvron et al., 2023 · LLaMA: Open and Efficient Foundation Language Models · arxiv.org/abs/2302.13971

14 DPO Direct Preference Optimization

TRAINING

WHAT IT INTRODUCED

Align a model to preference pairs (a chosen and a rejected answer) with one classification-style loss. No separate reward model, no RL loop.

WHY IT STILL MATTERS

Far simpler and more stable than RLHF with PPO, and as good or better in the paper's tests, so it became a go-to alignment step for open models.

One loss on (chosen, rejected) pairs · no reward model

SAY IT "DPO gets RLHF's result with a simple loss on preference pairs: no reward model, no RL."

Rafailov et al., 2023 · Direct Preference Optimization: Your Language Model is Secretly a Reward Model · arxiv.org/abs/2305.18290

15 DeepSeek-R1 Reasoning through RL

TRAINING

WHAT IT INTRODUCED

Reasoning trained with large-scale reinforcement learning on tasks with checkable answers, like maths and code, instead of copying human-written reasoning.

WHY IT STILL MATTERS

Self-reflection and verification emerged without being taught, the model was released openly, and its reasoning was passed down to smaller models.

No human-labelled reasoning steps: pure RL

SAY IT "Reward the right answer, not the right-looking steps, and reasoning emerges."

DeepSeek-AI, 2025 · DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning · arxiv.org/abs/2501.12948

See the thread? 06 shows reasoning can be prompted, 11 and 14 teach models what people prefer, and 15 lets models learn reasoning by themselves. That's the path from GPT-3 to today's reasoning models.

Where to start.

PICK A PATH

START HERE

How a chatbot with tools is built

[01 Transformer](#) → [03 GPT-3](#) → [11 RLHF](#) → [08 RAG](#) → [12 ReAct](#)**FINE-TUNING**

Adapting a model to you

[04 LoRA](#) → [14 DPO](#)**REASONING**

How models learned to think

[06 Chain-of-Thought](#) → [15 DeepSeek-R1](#)**MODEL DESIGN**

Inside the architecture

[09 BERT](#) → [10 MoE](#) → [13 LLaMA](#)**IMAGES**

Generating pictures

[02 DDPM](#) → [07 CLIP](#) → [05 Stable Diffusion](#)

Read a paper in 20 minutes

1. **Abstract:** what problem, what idea, what result.
2. **Figure 1:** most papers put the whole idea in one diagram.
3. **Conclusion:** what the authors think they proved.
4. **Results tables:** compared with what, and by how much.
5. **Method last,** and only the parts you still don't get.

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