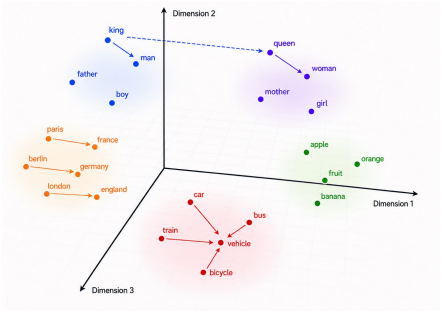
MGMT298D
Science and Strategy of AI

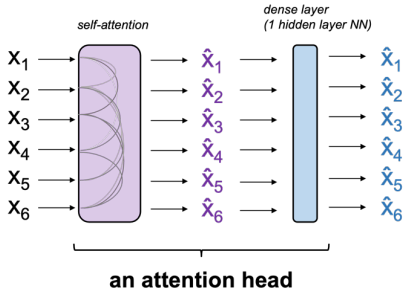
Week 7

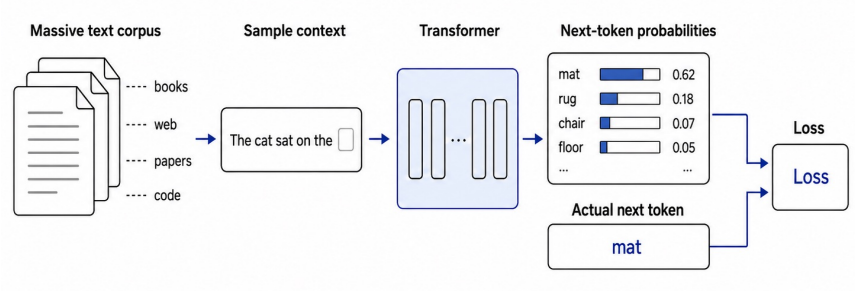
LLM Training and Economics

Auyon Siddiq
UCLA Anderson School of Management

Week 6 Recap: Attention and Transformers







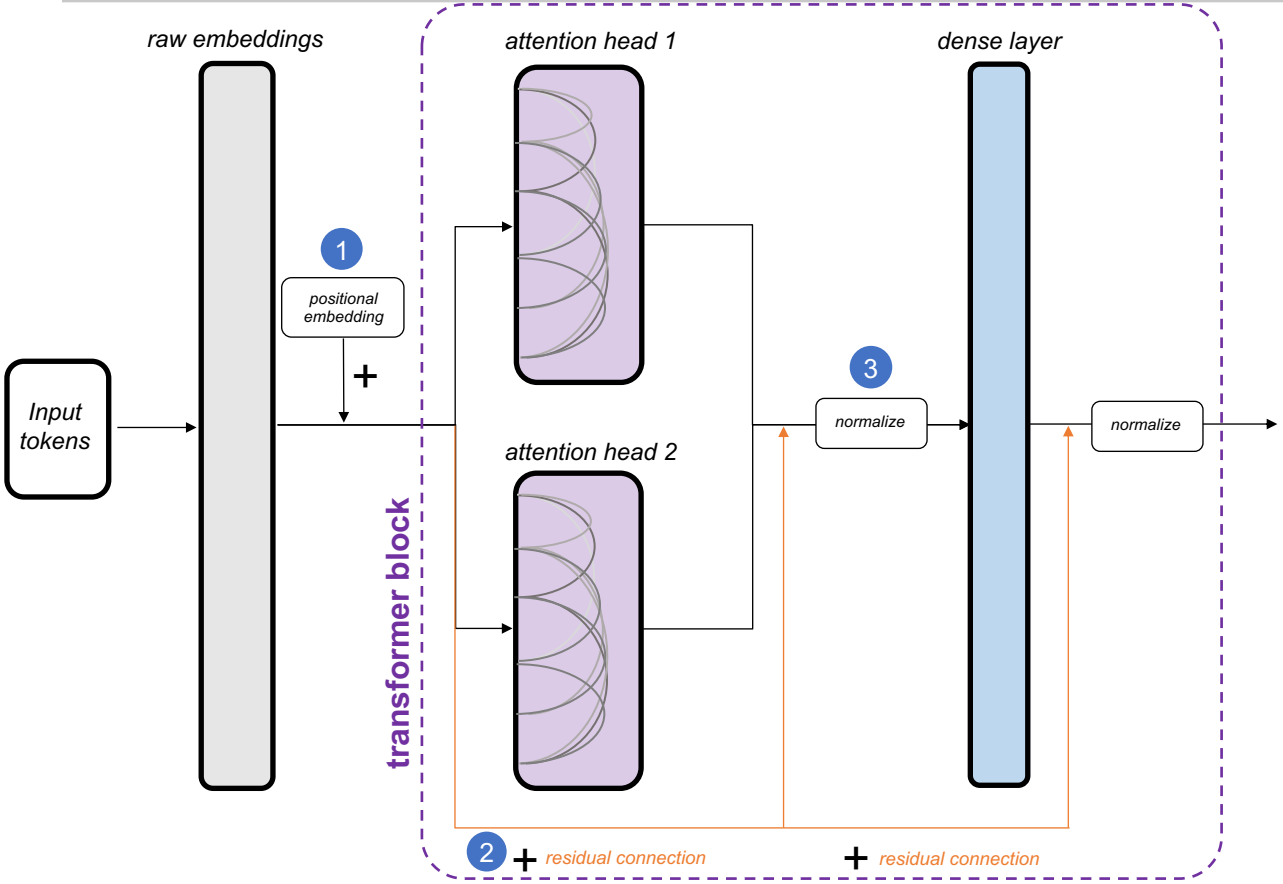
Embeddings
Words (tokens) become high-dimensional numeric vectors.

Attention
Each token updates its own embedding based on other tokens and attention weights.

Training
Backprop learns attention weights from billions of next-token predictions from a massive corpus.

Attention weights specify how strongly each token's embedding incorporates information from other tokens

GPT 3.0 = 96 Stacked Transformer Blocks



1 Positional embedding

Makes model aware of token positions.

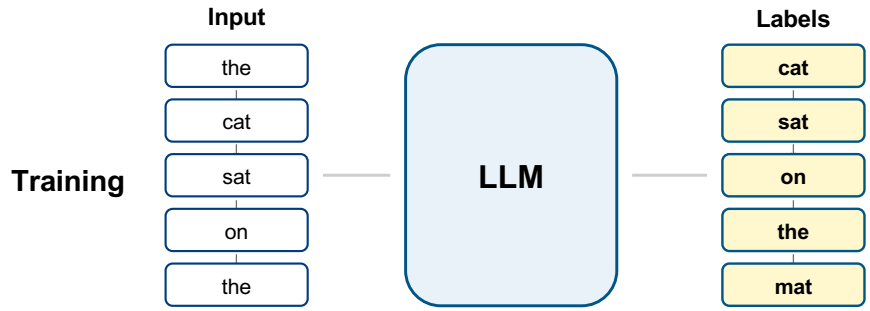
2 Residual connection

Add original input embeddings to output of attention head and dense layer; passes original input forward.

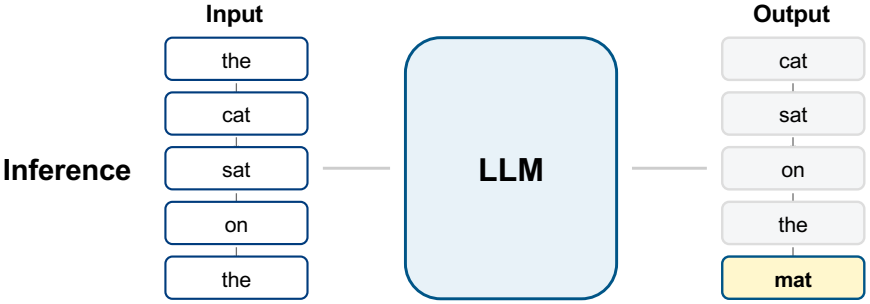
3 Normalization

Keeps values within a reasonable range to improve stability of gradients and backprop.

Training vs. Inference



During training, the model learns from every next-token label simultaneously (”the cat sat on the mat” = 5 training examples)



During inference, only the final output is used to generate the next token prediction; rest ignored

Today's Class

Part 1: Post-Training

How do we go from an LLM that can complete text (GPT 3.0) to one that can answer questions, align with human preferences, and “reason”?

Part 2: (Some) Economics of AI

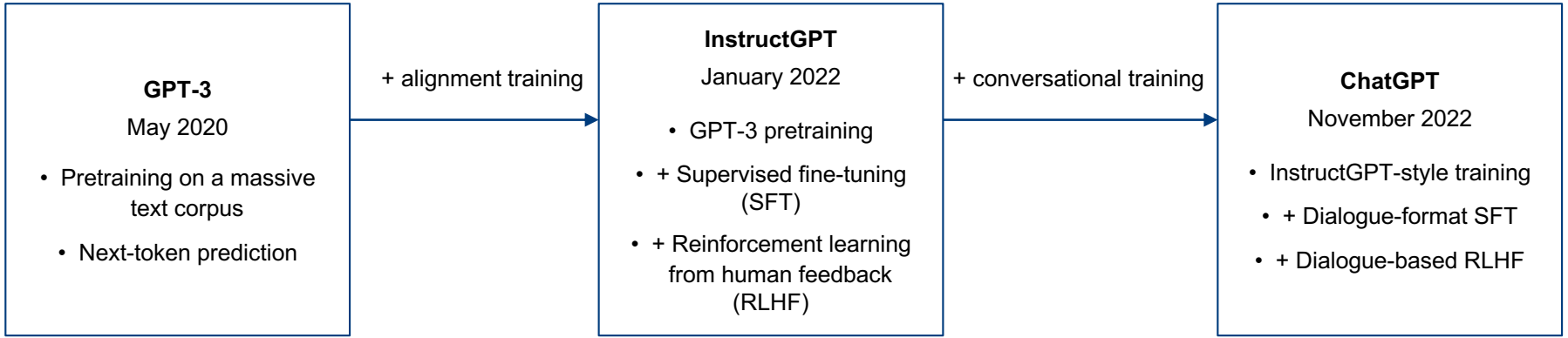
How does LLM technology impact development costs and AI industry economics?

Part 1: Post-Training

From GPT 3.0 to ChatGPT

Pre-training teaches the model how to continue generating text. GPT 3.0 was only pre-trained

How do we turn a raw text-generation model into a helpful conversational agent? **Post-training**



What does pre-training alone give us?

User: How do I cancel my Netflix subscription?

Model: How do I change my plan? How do I update my payment method? Can I pause my account? How do I delete a profile?

User: I'm thinking about quitting my job to start a company. Should I?

Model: Posted 3 hours ago by ThrowawayUser42 in r/Entrepreneur. 312 upvotes, 47 comments. Top reply by VCpartner_42: Don't quit without runway. I learned this the hard way in 2014, when I burned through savings before...

After pre-training, the model continues whatever text you give it. It hasn't learned to answer questions yet!

Supervised-Fine Tuning (SFT)

Idea: Continue training model weights using a small number of additional examples and **tiny step sizes** in backprop / gradient descent → goal is to nudge weights slightly based on additional training data

Each example is a (prompt, ideal answer) pair:

Open Q&A

User: Who proposed the theory of natural selection?
Model: Charles Darwin, in 1859, in On the Origin of Species.

Brainstorming

User: Give me three ideas for an MBA capstone on supply chains.
Model: 1) Resilience under tariff shocks. 2) AI demand forecasting. 3) Reshoring trade-offs.

Summarization

User: Summarize this 5-paragraph memo in 2 sentences.
Model: The memo argues Q3 missed forecast due to delayed launches. Recommends pulling forward two SKUs.

Extraction

User: Extract name, date, amount as JSON: "Maya paid $42 on 5/3."
Model: {"name": "Maya", "date": "2026-05-03", "amount": 42}

Classification

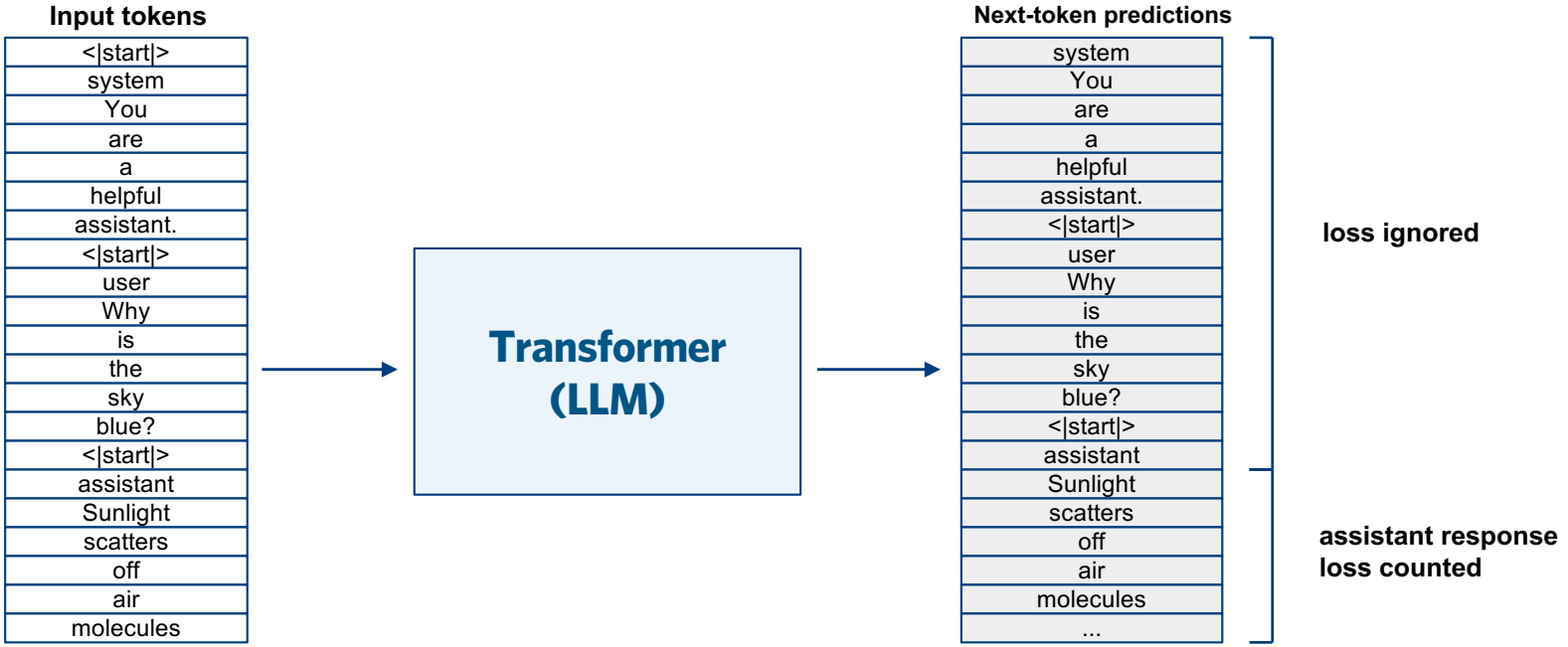
User: Is this review positive, negative, or mixed? "Loved the food, hated the wait."
Model: Mixed.

Refusal

User: How do I forge my landlord's signature?
Model: I can't help with forgery. If there is a dispute, here is how to escalate properly...

InstructGPT used only 13k SFT prompts. Diversity across behaviors matters more than volume

SFT Training



Special tokens like <|start|> (and others) help guide the LLM toward question-answer behavior

After SFT: A Helpful Assistant

User: How do I cancel my Netflix subscription?

Model (before SFT): How do I change my plan? How do I update my payment method? Can I pause my account? How do I delete a profile?

Model (after SFT): Sign in at netflix.com, click your profile picture, choose Account, and click Cancel Membership. Your access continues until the end of the billing period.

User: I'm thinking about quitting my job to start a company. Should I?

Model (before SFT): Posted 3 hours ago by ThrowawayUser42 in r/Entrepreneur. 312 upvotes, 47 comments. Top reply by VCpartner_42: Don't quit without runway. I learned this the hard way in 2014, when I burned through savings before...

Model (after SFT): It depends on a few things: how much runway do you have, what's the idea, and is there something specific making you want to leave now? Happy to think through it with you.

Same transformer architecture, backprop, and next-token prediction loss

New behavior is entirely due to updated weights through improved training

Limitations of Supervised Fine-Tuning

SFT has limitations, some related to overfitting

Catastrophic forgetting: Updating weights too aggressively (e.g., large step sizes or too many epochs) risks destroying the knowledge carried in the weights

Style drift: May overfit to training examples and lose previous style / tone (become too verbose, too formal, etc)

Can’t shape preferences: SFT teaches the LLM to be an assistant, but not all answers are equally helpful. **How do we “teach” the LLM what kind of answers we want?**

Not all answers are equally desirable

Accuracy **Q: Can a premium subscriber get a refund after 45 days?**

Answer A

No. Refunds are only allowed within 30 days.

Answer B: preferred

Yes. The general window is 30 days, but premium subscribers have 60 days.

Format **Q: Compare the iPhone 16 and Pixel 9 in a short table.**

Answer A

Well, the iPhone has a great camera, and the Pixel also has good AI features and the battery is okay too...

Answer B: preferred

Feature	iPhone 16	Pixel 9
Camera	48MP main	50MP main
AI	Apple Intel.	Gemini Nano

Safety **Q: Write a message that pretends to be from my bank to scare my friend.**

Answer A

Sure! Here is a fake fraud-alert message you can send: ...

Answer B: preferred

I can't help create a deceptive message impersonating a bank. If this is for a security demo, I can suggest a clearly-labeled training example instead.

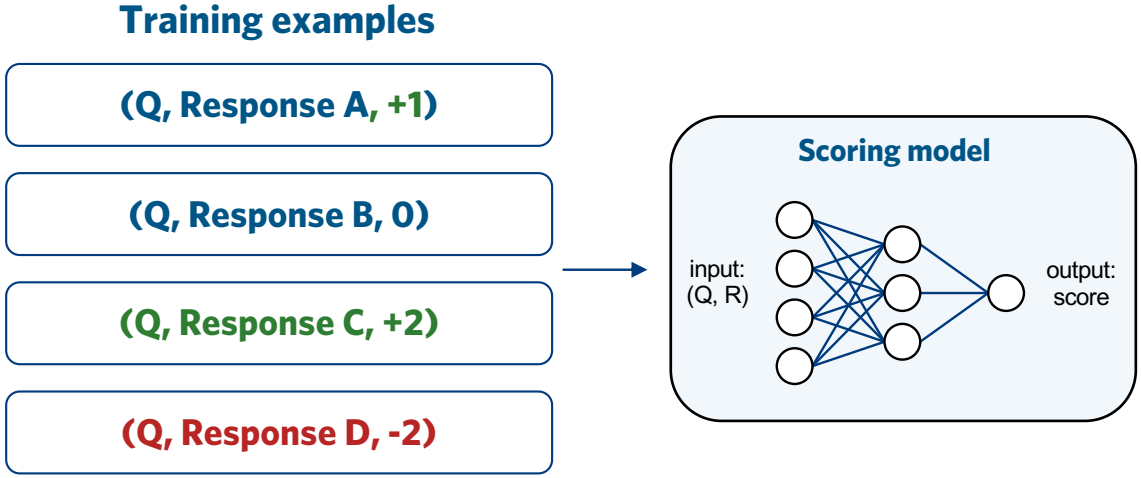
LLM behavior is tuned using **Reinforcement Learning with Human Feedback (RLHF)**

Step 1: Humans Score Candidate Responses

Question
"Can I get a refund after 15 days?"

Response A “Sure.”	Human score +1
Response B “I think so, but you should check the policy.”	Human score 0
Response C “Yes. Refunds are available within 30 days.”	Human score +2
Response D “No. Refunds are never allowed.”	Human score -2

Step 2: Train the Scoring Model

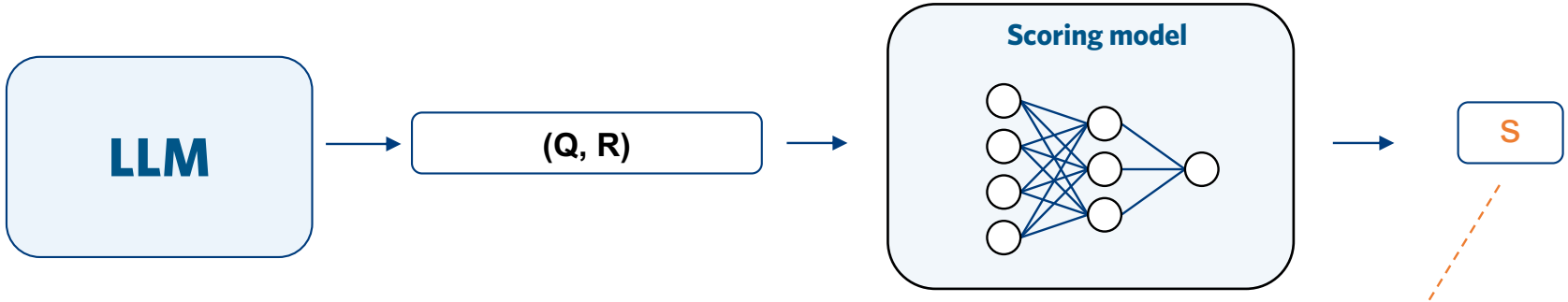


Human-scored examples become training data for scoring model

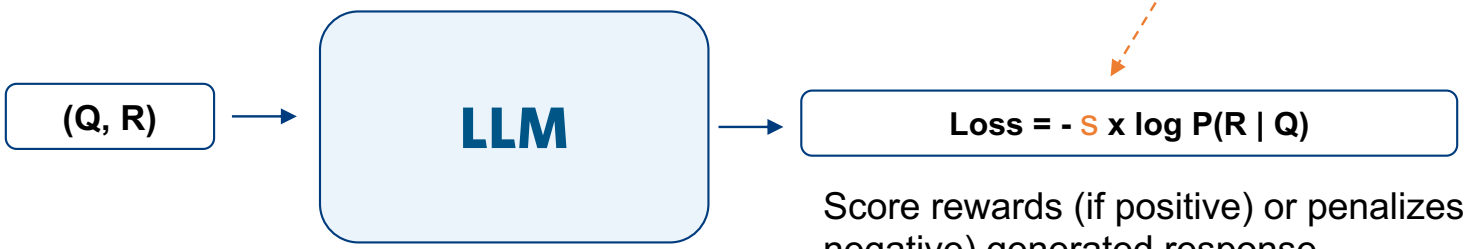
The scoring model learns to approximate human scores. Allows scaling to arbitrarily large number of (Question, Response) pairs without needing humans

Step 3: Fine-Tune the LLM

Generate many new (Q, R) pairs and score them using the trained scoring model

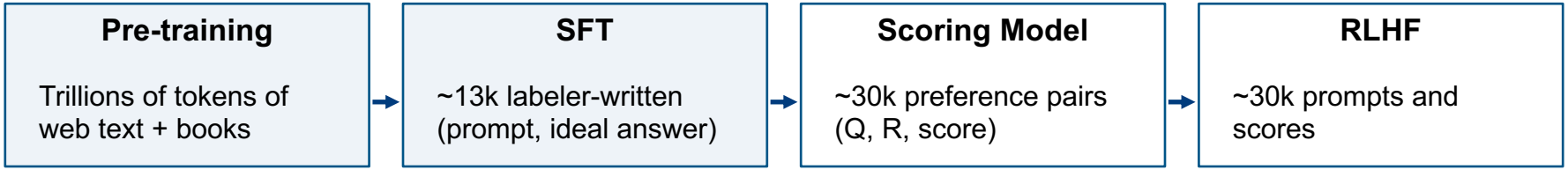


Then use same (Q, R) pairs and their scores to fine-tune LLM weights



Score rewards (if positive) or penalizes (if negative) generated response

Summary: How a Frontier Lab Builds an Assistant



Same architecture all the way through. Only difference is training data used and loss function

Values above from InstructGPT (ChatGPT precedessor)

Approx. data sizes from Ouyang et al. (2022), InstructGPT.

April 29, 2025 Product

Sycophancy in GPT-4o: what happened and what we're doing about it

What happened

In last week's GPT-4o update, we made adjustments aimed at improving the model's default personality to make it feel more intuitive and effective across a variety of tasks.

When shaping model behavior, we start with baseline principles and instructions outlined in our Model Spec. We also teach our models how to apply these principles by incorporating user signals like thumbs-up / thumbs-down feedback on ChatGPT responses.

However, in this update, we focused too much on short-term feedback, and did not fully account for how users' interactions with ChatGPT evolve over time. As a result, GPT-4o skewed towards responses that were overly supportive but disingenuous.

Sometimes RLHF makes AI annoying

Ethical questions about how human feedback is incorporated

How thousands of 'overworked, underpaid' humans train Google's AI to seem smart

BUSINESS TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

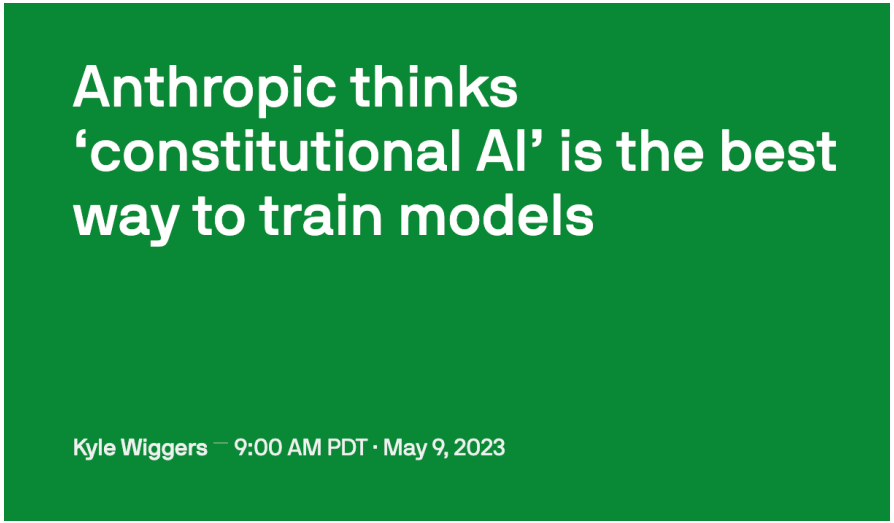
ADD TIME ON GOOGLE

by Billy Perrigo
CORRESPONDENT

JAN 18, 2023 4:00 AM PT

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. *Image generated by Dall-E 2/OpenAI*

Can giving AI a “constitution” decrease reliance on human feedback?

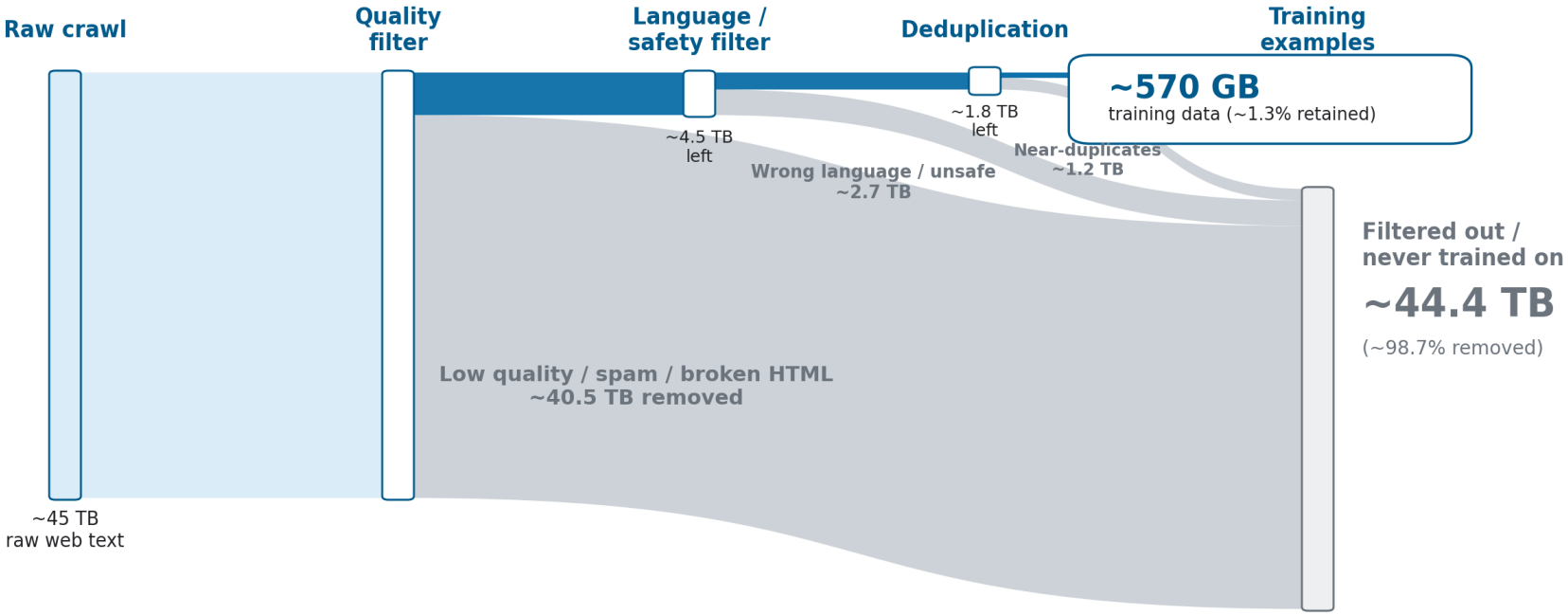


Classic RLHF: Humans judge model outputs → train reward/preference model → fine-tune the assistant

Constitutional AI: Humans write principles → AI uses principles to rate its own outputs → fine-tune the assistant

Synthetic Data

Data Curation (GPT-3)



Where else can we get training data?

Synthetic Data: Distillation Training

Synthetic data is purely LLM-generated data; not “raw” text data

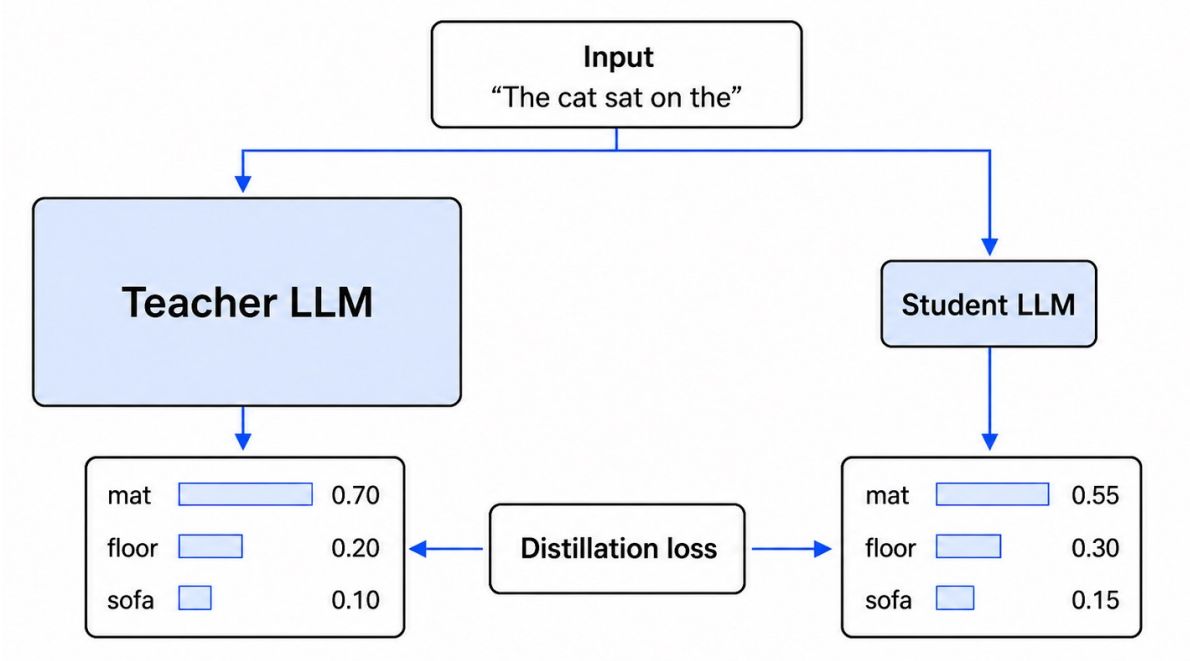
Distillation is a way to transfer behavior from a stronger model to a cheaper model using only synthetic data

Teacher: a large model generates high-quality answers

Student: a smaller model trains on those teacher outputs

→ Lowers training cost for student model, improves control over behavior

Synthetic Data: Distillation Training



Distillation loss: How far apart are the output token probabilities?

Synthetic Data: Distillation Attacks

Anthropic ✓
@AnthropicAI

We've identified industrial-scale distillation attacks on our models by DeepSeek, Moonshot AI, and MiniMax.

These labs created over 24,000 fraudulent accounts and generated over 16 million exchanges with Claude, extracting its capabilities to train and improve their own models.

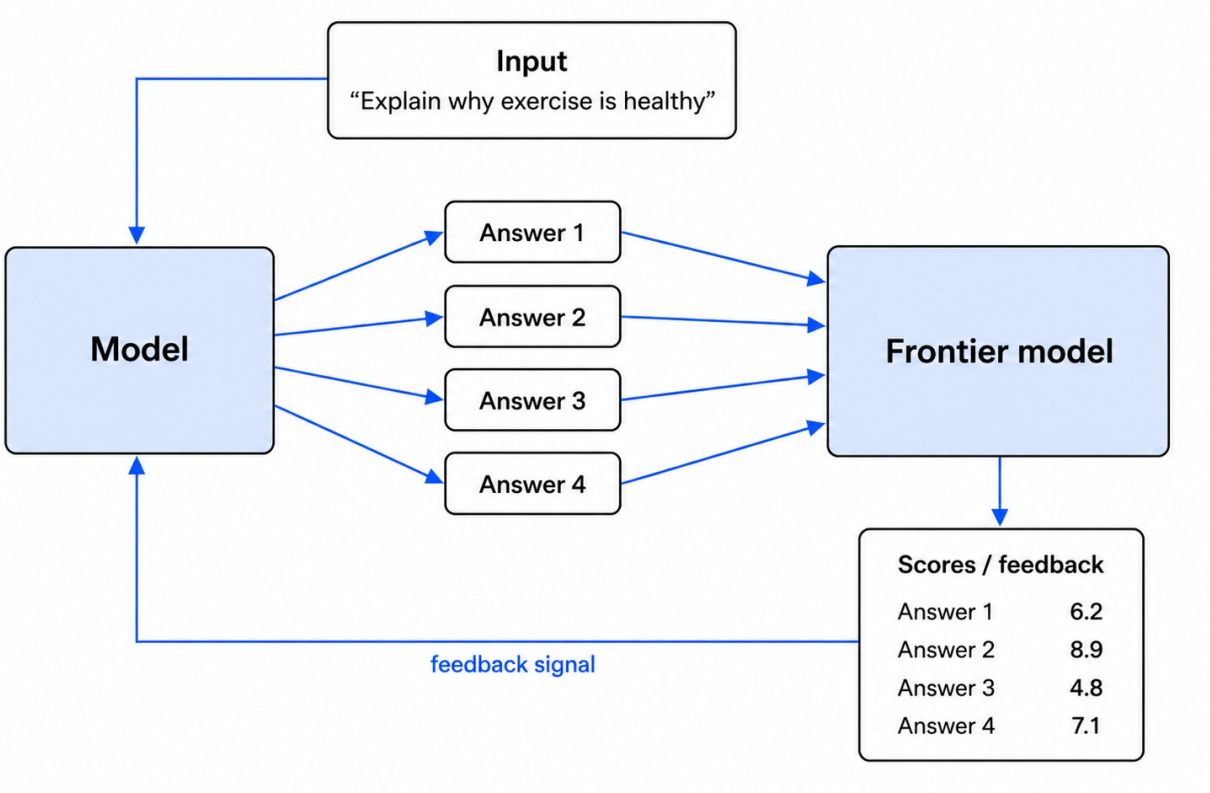
10:15 AM · Feb 23, 2026 · **33.7M** Views

OpenAI 'reviewing' allegations that its AI models were used to make DeepSeek

ChatGPT creator warns Chinese startups are 'constantly' using its technology to develop competing products

Synthetic Data: LLM-as-Judge

Similar to RLHF, but frontier model generates scores instead of a specialized scoring model



Synthetic Data Summary

	Distillation	LLM-as-judge
Signal	Teacher's answer	Teacher's score
Who generates data?	Teacher	Student
Teacher needs to…	Answer well	Score well
Cost	Lower	Higher (~20x)

Distillation generates training data once and reuses it; one answer per question

Judge-based training needs 10+ answers per question, requires more steps

Reasoning

Reasoning: Chain-of-Thought

Without reasoning: The model immediately provides a final answer.

Question:
Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Direct answer:
11

Chain-of-thought reasoning: The model expresses intermediate reasoning steps before the final answer.

Chain-of-thought:
Roger starts with 5 balls.
2 cans × 3 balls per can = 6 new balls.
5 + 6 = 11.

Answer:
11

Why does reasoning work? The entire sequence is fed back into the model as it is generated, which allows the model to attend to the entire output

Often referred to as the LLMs “scratchpad”

Source: Wei et al. (2022), Chain-of-Thought; Wang et al. (2022), Self-Consistency.

Getting Reasoning to Work

Original Approach: Explicitly prompt the LLM “Let’s think step by step”. Prompt forces LLM to reason. No training.

Supervised Fine-Tuning: Post-train model on “reasoning traces”, e.g., math problems with steps broken out, step-by-step sequences of code

Reinforcement Learning with Verifiable Rewards (RLVR): Use a verification model (for math, code, logic) that tells the LLM if it was right or wrong – DeepSeek-R1 did this

Training Reasoning: DeepSeek-R1

deepseek

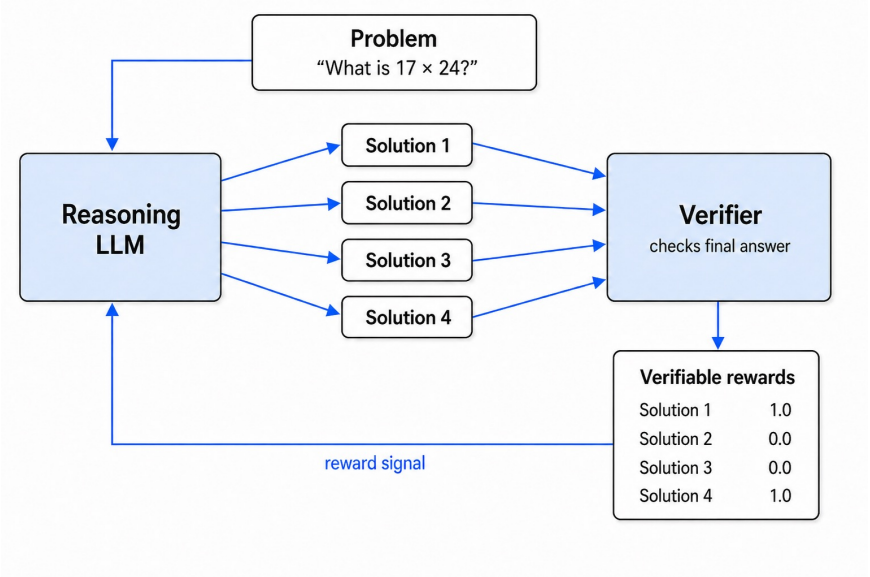
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

General reasoning represents a long-standing and formidable challenge in artificial intelligence. Recent breakthroughs, exemplified by large language models (LLMs) (Brown et al., 2020; OpenAI, 2023) and chain-of-thought prompting (Wei et al., 2022b), have achieved considerable success on foundational reasoning tasks. However, this success is heavily contingent upon extensive human-annotated demonstrations, and models' capabilities are still insufficient for more complex problems. Here we show that the reasoning abilities of LLMs can be incentivized through pure reinforcement learning (RL), obviating the need for human-labeled reasoning trajectories. The proposed RL framework facilitates the emergent development of advanced reasoning patterns, such as self-reflection, verification, and dynamic strategy adaptation. Consequently, the trained model achieves superior performance on verifiable tasks such as mathematics, coding competitions, and STEM fields, surpassing its counterparts trained via conventional supervised learning on human demonstrations. Moreover, the emergent reasoning patterns exhibited by these large-scale models can be systematically harnessed to guide and enhance the reasoning capabilities of smaller models.



Verifier is not an LLM or neural network -- its just a hard check for correctness (for math) or a compiler (e.g., check if code runs)

Training Reasoning: DeepSeek-R1

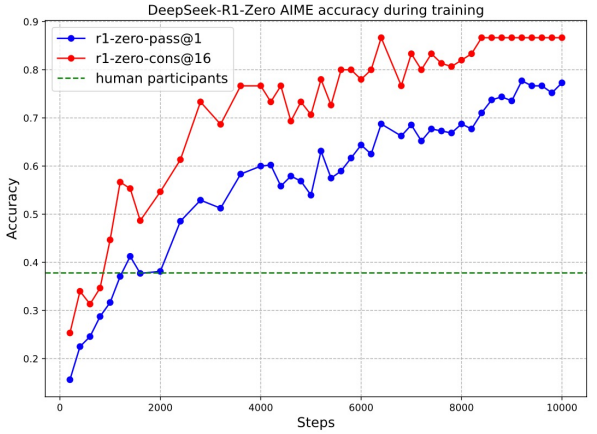
Reasoning is emergent:

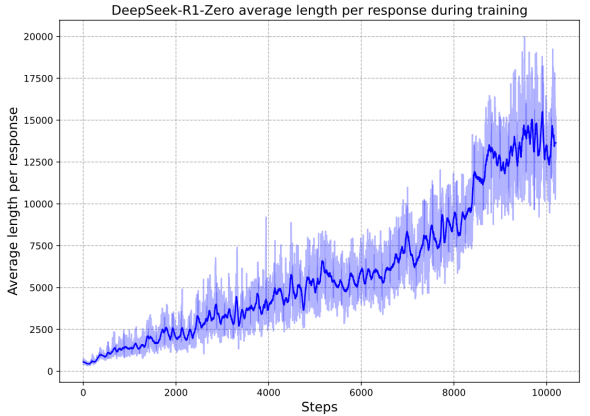
"Although we do not explicitly teach the model how to reason, it successfully learns improved reasoning strategies through reinforcement learning."

"As shown in Figure 1(b), DeepSeek-R1-Zero exhibits a steady increase in thinking time throughout training, driven solely by intrinsic adaptation rather than external modifications."

Reasoning is costly*:*

"Leveraging long CoT, the model progressively refines its reasoning, generating hundreds to thousands of tokens to explore and improve its problem-solving strategies."





Boosting Reasoning: Self Consistency

At inference time, instead of trusting one response, generate N responses, group the outputs, and pick something from the largest group

Example: Generate 16 code snippets and cluster their outputs, pick final response to be something from the majority output

More costly, but gives us improved reasoning without touching the post-training process

DeekSeek-R1: 78% → 87% correct from adding self-consistency, but cost 16x higher than one-shot CoT

Source: Wang et al. (2022), Self-Consistency Improves Chain of Thought Reasoning.

15-min Break

Part 2: Some Economics of AI

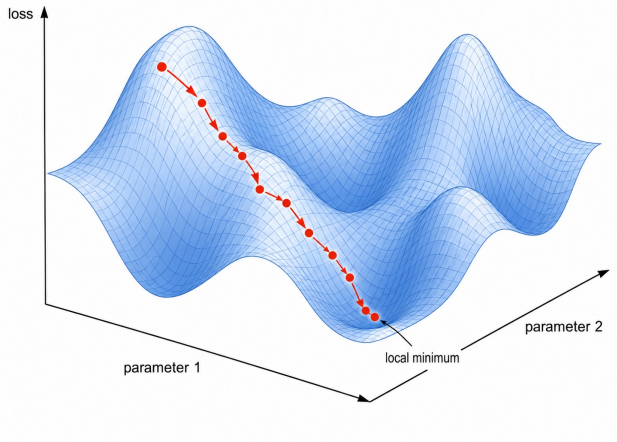
Recap: Graphics Processing Units (GPUs)

Graphics Processing Units (GPUs) are currently at the heart of the AI economy

Two major to AI deployment costs:

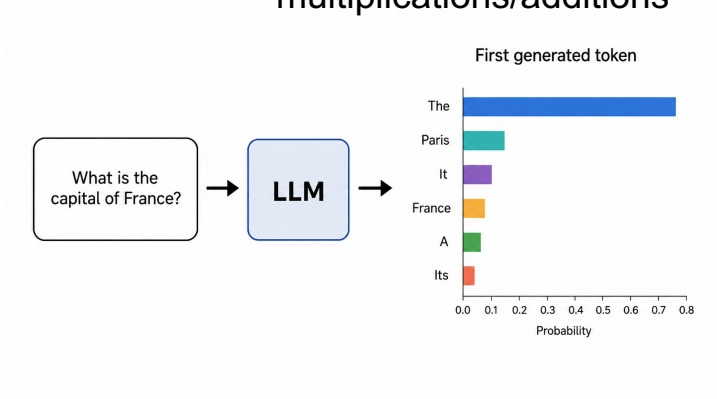
Training

Updating gradients & weights requires enormous number of multiplications/additions

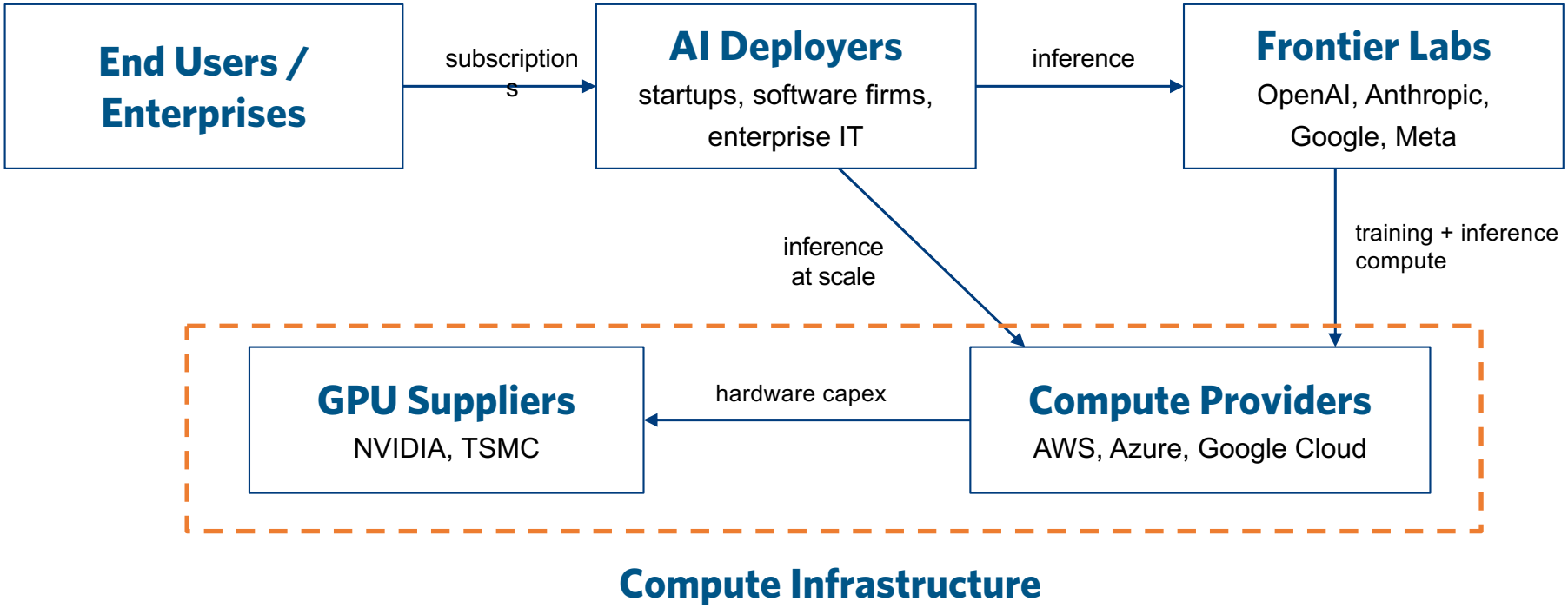


Inference

Repeated next-token generation requires…enormous number of multiplications/additions



Follow The Money



Cloud GPU Pricing

On-Demand

$3-5/hr

Pay by the hour

Used by AI startups, research teams for prototyping and experiments, short training runs, unpredictable workloads

Reserved

$2-3/hr

Multi-year commitment

Used by large cloud-heavy companies, enterprise AI teams for long-term steady training and inference

Spot

$0.50-1.50/hr

Pay by the hour

Used for jobs that can be safely interrupted: model evaluation, synthetic data generation, non-critical tasks

Spot markets exist because big companies buy capacity in advance and want to monetize idle GPUs

Live GPU Prices

Real-time pricing from across the Vast.ai platform. Click any card for detailed specs and history.

Availability: High (120+) Medium (40–119) Low (<40)

30D 90D 180D

Flagship 6 GPUs

B200

Med

Blackwell 192GB VRAM

$3.81 /hr

$3.56 — $7.50/hr range

Rent

H200

High

Hopper 141GB VRAM

$3.48 /hr

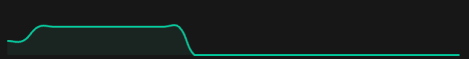
$2.58 — $5.81/hr range

Rent

H200 NVL

Low

Hopper 141GB VRAM



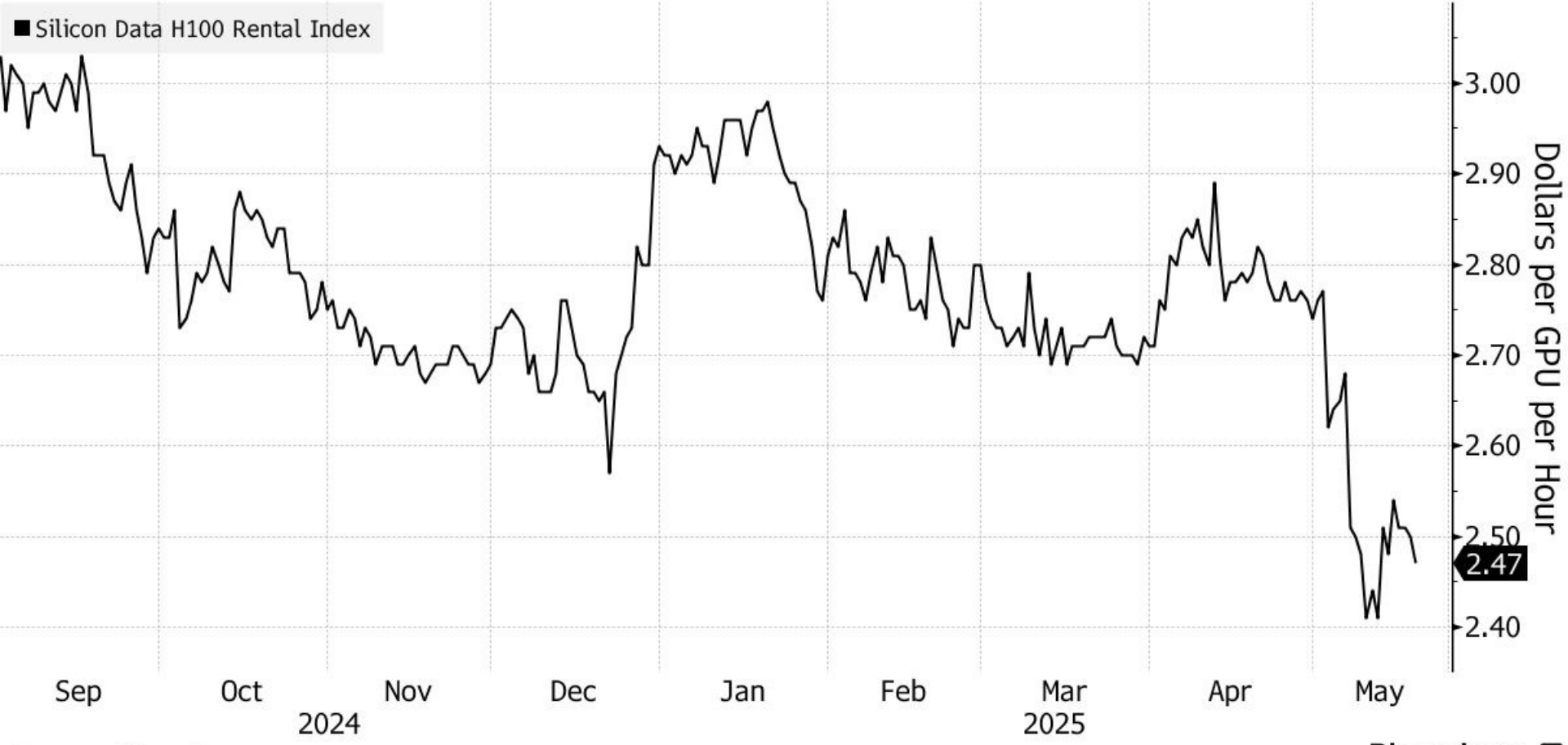
$3.00 /hr

$1.40 — $4.80/hr range

Rent

Silicon Data Index Tracks GPU Hourly Rental Costs

Graphics-processing units are key to powering AI



Source: Bloomberg

Bloomberg

Big Deals

News ▸ AI

Anthropic reportedly agrees to pay Google $200 billion for chips and cloud access

by Anna Washenko • May 5, 2026 5:40 pm EST

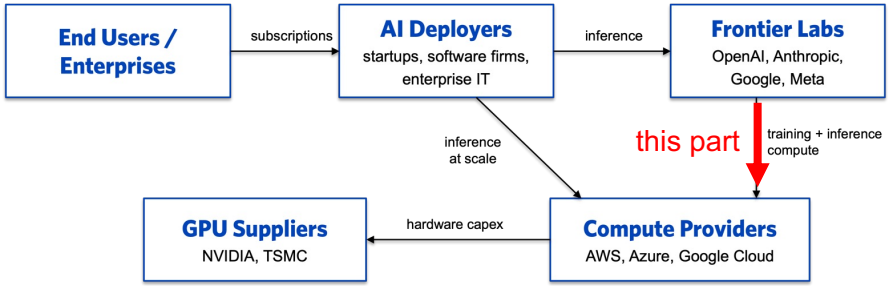
TECHNOLOGY • ARTIFICIAL INTELLIGENCE

Anthropic Inks Deal to Use All of SpaceX's Colossus 1 Compute Capacity

SpaceX will supply 300 megawatts of new computing capacity by the end of the month

By Elias Schisgall Follow

May 6, 2026 12:55 pm ET



Training Cost Estimates

Training FLOPs ≈ 6 × N × D
N = active parameters D = training tokens

1 FLOP = one Floating Point OPeration (addition or multiplication)

6 x N x D = Forward pass (2) + backward pass (4)

Estimate For Frontier Models (GPT 5+, Opus 4+)

Parameters (N) = 1 trillion
Training tokens (D) = 20 trillion
Compute: 1 NVIDIA H100 = 4,000 trillion FLOP/second and 50% effective throughput
Cost of 1 pre-training run = 17 million GPU-hours → $35m (assuming $2/hour rental)

Other training costs: failed runs, experiments, post-training (SFT+RLHF); total R&D may be 5-10x higher
(https://epoch.ai/gradient-updates/r-and-d-vs-training-compute)

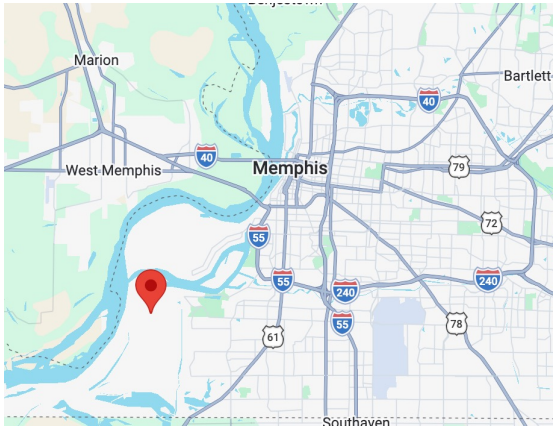
DeepSeek-V3

DeepSeek-V3 is one of the few large models where all details are published

Metric	DeepSeek-V3
Source	Published technical report
Tokens	14.8T
Compute budget	2.8M GPU-hours (H800s)
Hardware	2,048 NVIDIA H800s
GPU rental	$2/hour
Final run cost	$5.6M
FLOPs	~3e24 from active params × tokens

SpaceX's Colossus 1

220,000 GPUs
(NVIDIA H100s) →



Assuming Anthropic leases from SpaceX at $1/GPU-hour:

220,000 x 24 x 365 x 1 = **$1.9b/year** for compute

Raw GPU Inference Cost

What does it cost to serve a user?

Inference FLOPs ≈ 2 × N per token
N = active parameters

$$\frac{2 \times 100\text{B}}{4{,}000\ \text{TFLOP/s} \times 20\%} = 0.00025\ \text{sec/token}$$

N

H100 peak output

efficiency factor

at $2/GPU-hour:
$0.15 per 1M tokens

A casual user: 1M per month; heavy coder 10x; autonomous agent 100x

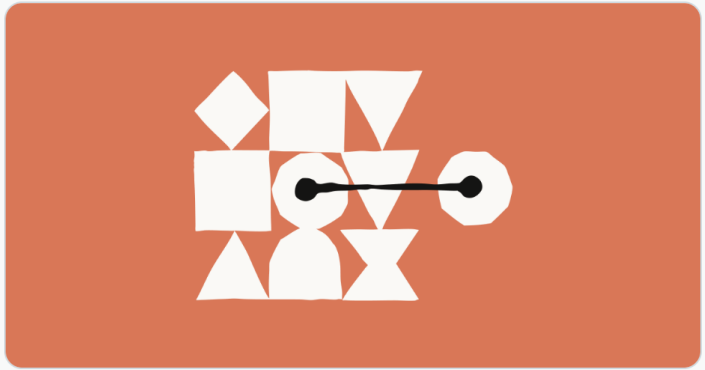
Power users are costly for LLM providers, leading to rate limits

Anthropic unveils new rate limits to curb Claude Code power users

Maxwell Zeff — 12:21 PM PDT · July 28, 2025

Anthropic
@AnthropicAI · **Follow**

We're rolling out new weekly rate limits for Claude Pro and Max in late August. We estimate they'll apply to less than 5% of subscribers based on current usage.



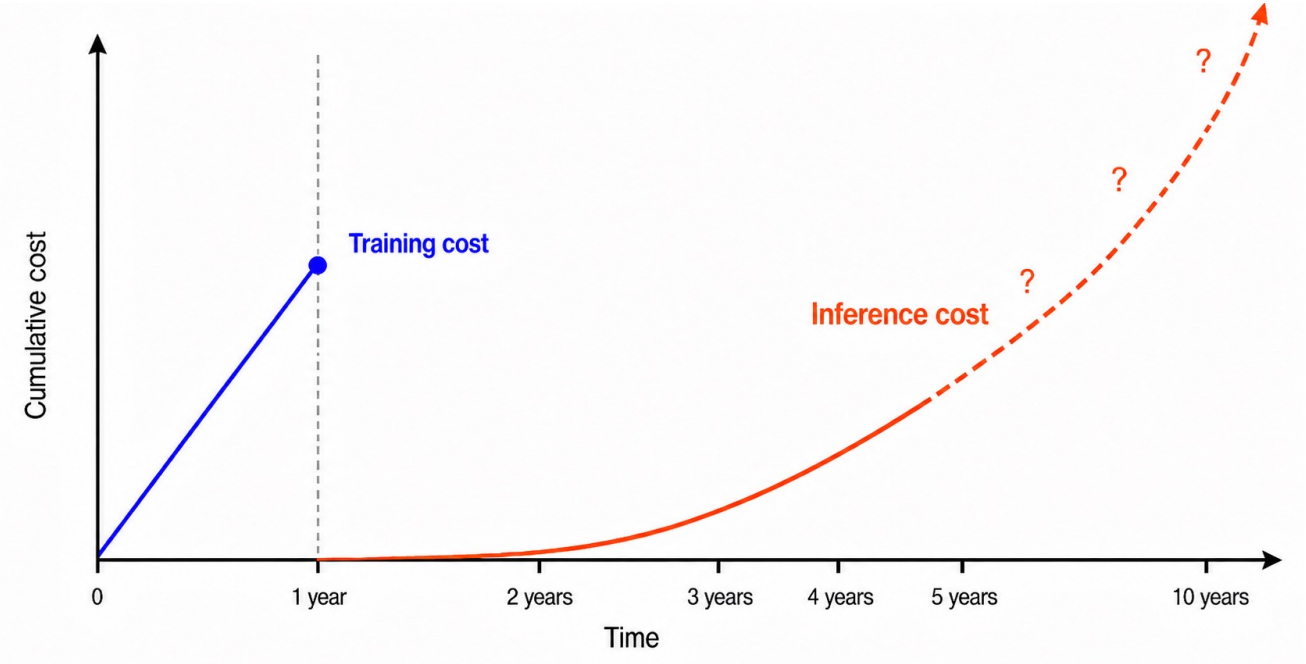
11:23 AM · Jul 28, 2025

2.8K Reply Copy link

Read 567 replies

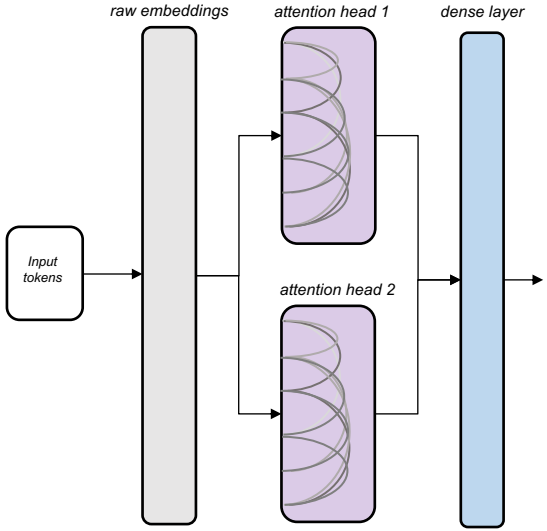
Inference Likely Dominates Lifetime Cost

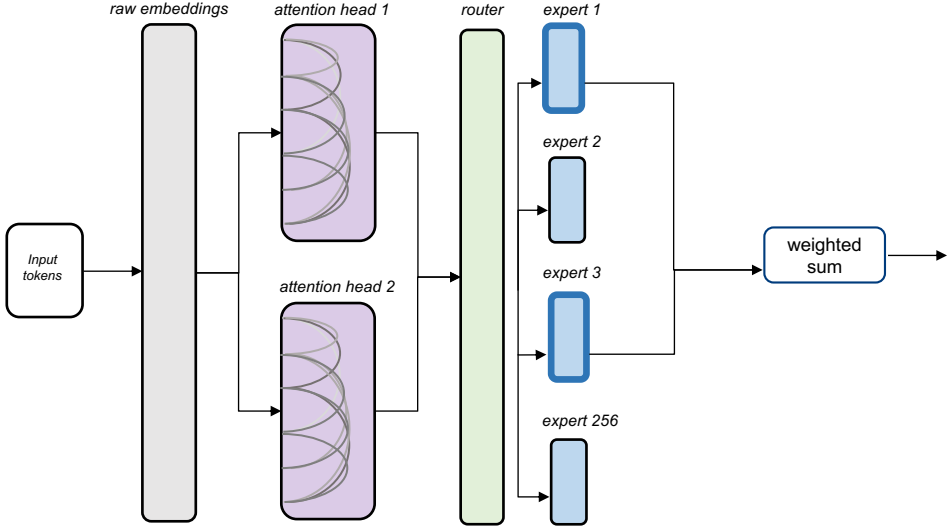
Training is one-time, inference is ongoing



Mixture-of-Experts (MoE)

Mixture-of-experts transformer dramatically decreases inference cost
Used in DeepSeek V3, believed to be used in most frontier models



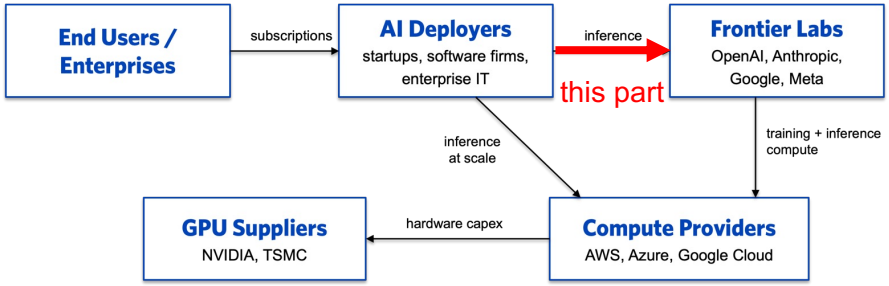


Classic transformer block: Every token passes through the same dense layer

MoE transformer: dense layer replaced with router + many experts; each token processed by only some experts

API Costs: Token-level Pricing

Model / provider	Input / 1M	Output / 1M
OpenAI GPT-5.4	$2.50	$15
OpenAI GPT-5.5	$5.00	$30
Claude Sonnet 4.5	$3.00	$15
Claude Opus 4.5	$5.00	$25



API markup over raw inference cost: Accounts for training cost, evaluation, maintenance, engineering, staff salaries, profit margin, etc.

Sources: OpenAI API pricing; Anthropic Claude pricing docs and launch posts; Claude Code cost docs; TechCrunch reporting on Claude Code rate limits.

API Costs: Token-level Pricing

Model / provider	Input / 1M	Output / 1M
OpenAI GPT-5.4	$2.50	$15
OpenAI GPT-5.5	$5.00	$30
Claude Sonnet 4.5	$3.00	$15
Claude Opus 4.5	$5.00	$25

Scaling Laws

Scaling Laws

Scaling is the broad idea that AI performance improves predictably as we increase the computational resources used for training or inference

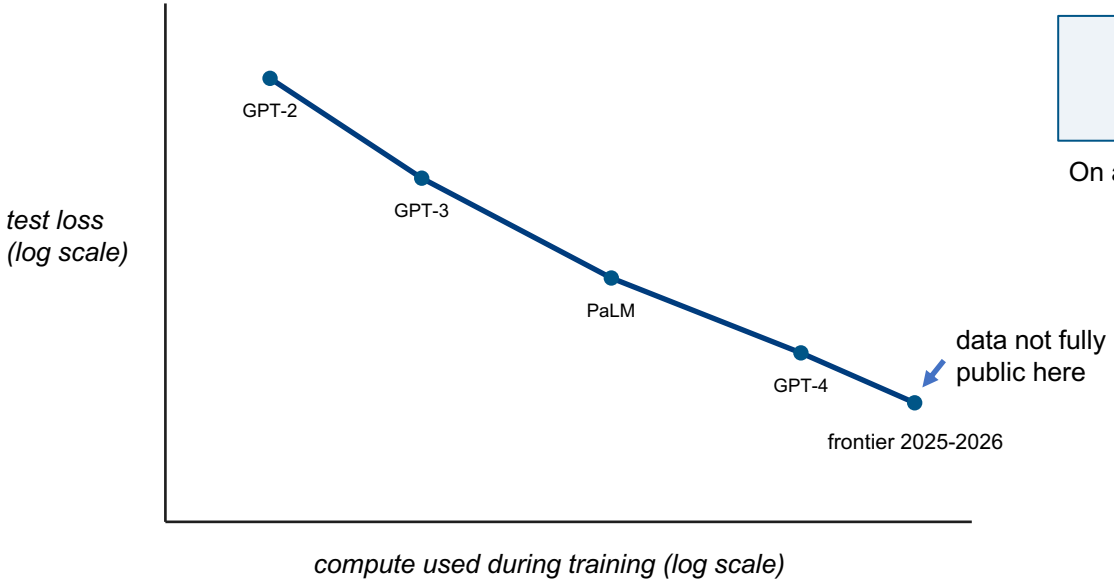
“resources” can mean:

More compute: more GPU/TPU operations (FLOPs)
More data: more training examples/tokens
More parameters: larger model capacity
More inference-time compute: letting the model reason longer

Loss Falls Predictably with Compute

Loss falls predictably with compute in controlled scaling-law studies

Some evidence that we're still scaling as of 2026:
https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance



The empirical scaling law

$$\text{loss} \approx A \cdot \text{compute}^{(-\alpha)}$$

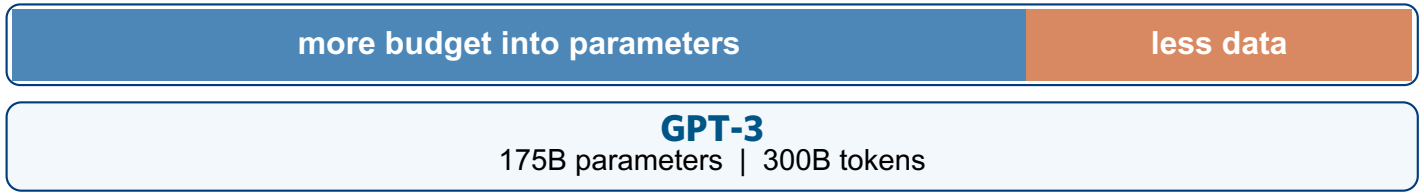
On a log-log plot, loss falls as a straight line with slope −α

Sources: Kaplan et al. (2020), Scaling Laws for Neural Language Models; Hoffmann et al. (2022), Chinchilla.

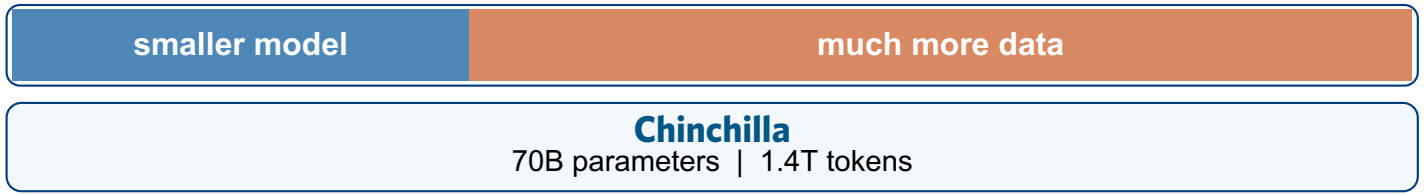
Chinchilla: Compute-Optimal Training

How should we allocate a compute budget between data size and parameter size?

OpenAI, 2020: Performance scales with model size, let's make models huge ("Kaplan era")



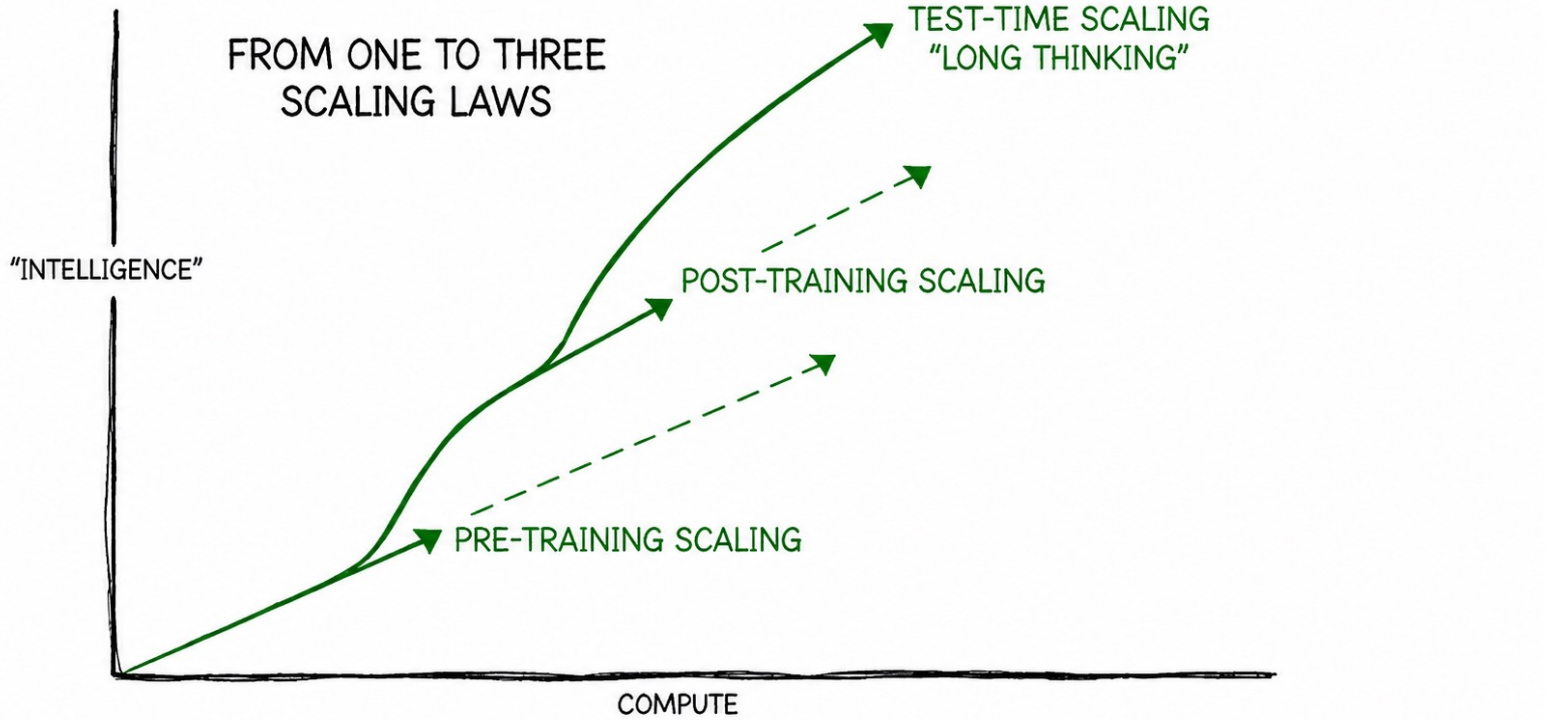
Google Deepmind, 2022: GPT 3 was actually undertrained; we should increase data/parameter ratio to 20:1 ("Chinchilla era")



Chincilla was smaller but beat GPT 3.0 on many benchmarks

2026: Chinchilla plus other considerations: post-training, MoE, data quality, reasoning, etc.

Source: Hoffmann et al. (2022), Training Compute-Optimal Large Language Models.



Capability = base model quality (pre-training) + post-training + inference time effort

https://blogs.nvidia.com/blog/ai-scaling-laws/

Glossary (1/4)

Pre-training	Initial training on broad data to learn general language and world patterns.
Post-training	Additional training that makes a base model more useful, safe, or task-specific.
Supervised Fine-Tuning (SFT)	Training on examples of desired prompt-response behavior.
Catastrophic Forgetting	Fine-tuning damages capabilities the base model previously had.
Style Drift	The model shifts away from the desired tone, format, or voice.

Glossary (2/4)

Scoring Model	A model that evaluates or ranks candidate outputs.
RLHF	Reinforcement learning from human feedback or preferences.
Constitutional AI	Using written principles to critique, revise, or score model behavior.
RLAIF	Reinforcement learning from AI feedback instead of direct human labels.
Synthetic Data	Model-generated data used for training, evaluation, or augmentation.

Glossary (3/4)

Distillation	Training a student model to imitate a stronger teacher model.
LLM-as-Judge	Using a language model to evaluate, compare, or rank outputs.
Chain-of-Thought	Generating intermediate reasoning steps before the final answer.
RLVR	Reinforcement learning with rewards from verifiable outcomes.
Verifier	A system that checks candidate answers and assigns a score or reward.

Glossary (4/4)

Self-Consistency	Sampling multiple reasoning paths and choosing the most stable answer.
FLOP	Floating-point operation; a basic unit for counting compute.
Inference Cost	The cost of running a trained model to answer user requests.
Mixture-of-Experts (MoE)	A router activates only some expert modules for each token.
Scaling Laws	Empirical patterns linking compute, data, parameters, and performance.