What it takes to train a large language model

Six steps turn raw text into a chatbot. Each one draws on five resources: data, compute, energy, time, and people. Pretraining burns most of the compute and power, but the bill is mostly chips and staff, and the biggest labs now share the least.

Data · Compute · Energy · Time · People

Bindu Bhargava Reddy Chintam
AIML-501 Model Development
Indiana Wesleyan University
Figures checked October 2026

The six steps, and what each one uses

The bars under each step show how heavily it draws on each resource: one segment is light use, three is the main cost of that step. The shaded step is where most of the money and power goes.

1 Collect data

Crawl the public web, buy or license text, add books, code, and human-written examples.

Llama 3 learned from 15.6 trillion tokens. Reddit reportedly licenses its posts to Google for about $60 million a year.

Llama Team, 2024; Tong et al., 2024

2 Clean and tokenize

Remove duplicates, spam, and harmful or private text, then split the rest into tokens.

GPT-3's Common Crawl slice shrank from 45 TB to 570 GB after filtering.

Brown et al., 2020

3 Pretrain

Thousands of GPUs or TPUs learn to predict the next token, for weeks or months without stopping.

Llama 3.1 405B used 3.8×10^{25} operations on up to 16,384 GPUs and had 466 interruptions in one 54-day stretch.

Llama Team, 2024

4 Fine-tune and align

Train on human-written answers, then on human or AI rankings (RLHF, Constitutional AI).

InstructGPT relied on about 40 hired labelers. DeepSeek-V3 spent 5K GPU-hours here versus 2,664K on pretraining.

Ouyang et al., 2022; DeepSeek-AI, 2024

5 Test and red-team

Run benchmarks and safety tests, invite outside experts to break it, and fix what they find.

GPT-4 finished training in August 2022, then spent about six months in safety work before its March 2023 launch.

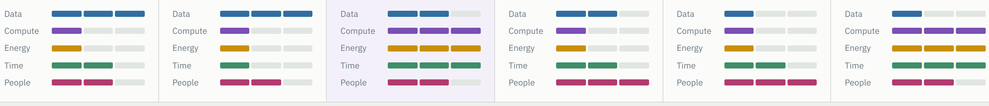
OpenAI, 2023

6 Deploy and serve

Run the model for millions of users, monitor it, and update it. This cost never stops.

Data centers used about 415 TWh in 2024, 1.5% of world electricity, and may reach 945 TWh by 2030.

IEA, 2025



The steps loop. Problems found in testing or in use send teams back to the data or to fine-tuning, so a model is never really finished after step 6.

Bar levels are my judgment, based on the examples shown and the sources in the footer.

Five resources and how each one drives cost

The assignment names data, compute, energy, and time. I added people because research staff are one of the two largest costs, even though they never appear in a GPU count.

Resource	What drives it	How it is measured, with an example	Share of a frontier model's cost
Data Tokens of text and code	Model size sets the appetite: Chinchilla research suggests about 20 tokens per parameter. Licensing, filtering, and labeling add cost.	GPT-3 learned from 300 billion tokens; Llama 3.1 405B from 15.6 trillion, 52 times more.	**Varies** Licensing and labeling; not broken out in public studies
Compute Operations and GPU-hours	Compute ≈ 6 × parameters × tokens, so bigger models trained on more data grow it fast.	Llama 3.1 405B: 30.8 million H100 GPU-hours. GPT-4's training hardware is estimated to have cost $800M to buy.	**47–64%** Chips, servers, and networking, spread over their working life
Energy Megawatt-hours and carbon	GPU-hours × power per chip × data center overhead. The local grid decides the carbon.	GPT-3 used about 1,287 MWh. BLOOM used 433 MWh on France's low-carbon grid and emitted about 25 t of CO_2 against GPT-3's 552 t.	**2–6%** Small in dollars, large in emissions
Time Days on the cluster, months overall	Cluster size, hardware failures, and safety testing. More chips shorten the run but raise the cost.	BLOOM trained for 118 days. DeepSeek-V3 pretrained in under two months on 2,048 GPUs.	**Indirect** Time multiplies every other cost
People Researchers, engineers, labelers	Bigger projects need larger teams: GPT-3 listed 25 contributors, GPT-4 listed 284, Gemini 941.	Human feedback, red-teaming, and safety reviews all depend on people.	**29–49%** R&D staff including equity

Eight models, side by side

Every number carries a tag that says where it came from. Gaps are real: the companies behind GPT-4, Gemini, and Claude have not published these figures.

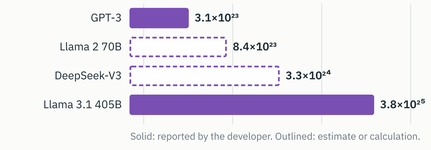
Model	Parameters	Training tokens	Compute (operations)	GPU-hours	Energy (MWh)	CO_2e (t)	Cost ($M)
GPT-3 OpenAI, 2020	175B [R]	300B [R]	3.1×10^{23} [R]	3.6M [E]	1,287 [E]	552 [E]	n/d
BLOOM BigScience, 2022	176B [R]	n/d	n/d	1.08M [R]	433 [R]	25 [R]	n/d
Llama 2 70B Meta, 2023	70B [R]	2T [R]	8.4×10^{23} [C]	1.72M [R]	688 [C]	291 [R]	n/d
GPT-4 OpenAI, 2023	n/d	n/d	n/d	n/d	n/d	n/d	40–78 [E]
Gemini Ultra Google, 2023	n/d	n/d	n/d	n/d	n/d	n/d	30–191 [E]
Claude 3 Anthropic, 2024	n/d	n/d	n/d	n/d	n/d	n/d	~100† [S]
Llama 3.1 405B Meta, 2024	405B [R]	15.6T [R]	3.8×10^{25} [R]	30.8M [R]	21,600 [C]	8,930 [R]	n/d
DeepSeek-V3 DeepSeek, 2024	671B* [R]	14.8T [R]	3.3×10^{24} [C]	2.79M [R]	n/d	n/d	5.6 [R]

*DeepSeek-V3 is a mixture-of-experts model: 671B parameters in total, 37B used per token. Energy marked C assumes every GPU ran at peak power, so it is an upper bound. †Anthropic's CEO said in mid-2024 that models then in use cost about $100M to train; Claude 3's own cost is not disclosed.

[R] Reported by the developer · [E] Estimated by researchers · [C] My calculation from reported numbers · [S] Executive's public statement · n/d Not disclosed

Training compute

Total operations, log scale

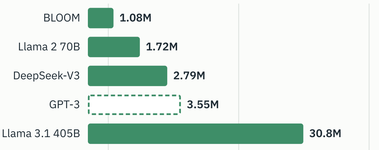


Solid: reported by the developer. Outlined: estimate or calculation.

R = reported; C = 6 × parameters × tokens (active parameters for DeepSeek-V3).

GPU-hours

Accelerator hours for the final run, log scale

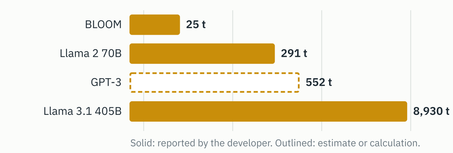


Solid: reported by the developer. Outlined: estimate or calculation.

GPT-3 assumes 10,000 V100 GPUs for 14.8 days (Patterson et al., 2021).

Carbon emitted in training

Tonnes of CO_2 equivalent, log scale

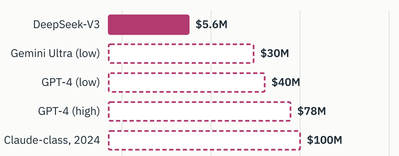


Solid: reported by the developer. Outlined: estimate or calculation.

Llama 3.1 is location-based. BLOOM is the energy-only figure from Luccioni et al. (2023).

Training cost

Millions of US dollars, log scale



Solid: reported by the developer. Outlined: estimate or calculation.

Low = amortized hardware and energy (Cottier et al.); high = cloud-rental price (AI Index). DeepSeek's figure excludes research runs.

What the numbers say

121×
more compute for Llama 3.1 405B than for GPT-3, four years apart. Compute is roughly 6 × parameters × training tokens, and both grew.

2–6%
of a frontier model's cost is energy. Chips and servers (47–64%) and staff (29–49%) dominate, yet Llama 3.1 still emitted about 16 times GPT-3's carbon.

0.2%
of DeepSeek-V3's GPU time went to post-training. Alignment is cheap in compute but heavy in human work: labelers, red-teamers, and months of testing.

$40M or $78M?
Two careful estimates of GPT-4's cost differ by about two times because they count different things. Without disclosure, every frontier number is a range.

Sources. Anthropic (2024) Claude 3 model card; Brown et al. (2020); Cottier et al. (2025); DeepSeek-AI (2024); Fitzmaurice (2024); Gemini Team (2023); Hoffmann et al. (2022); IEA (2025); Kaplan et al. (2020); Llama Team (2024) and Meta (2024) Llama 3.1 model card; Luccioni et al. (2023); Maslej et al. (2024) AI Index; OpenAI (2023); Ouyang et al. (2022); Patterson et al. (2021); Tong et al. (2024); Touvron et al. (2023). Full APA 7 references are in the accompanying document. Energy marked C is GPU-hours times peak chip power, so it is an upper bound. Llama 3.1 carbon is location-based; Meta reports zero market-based emissions.