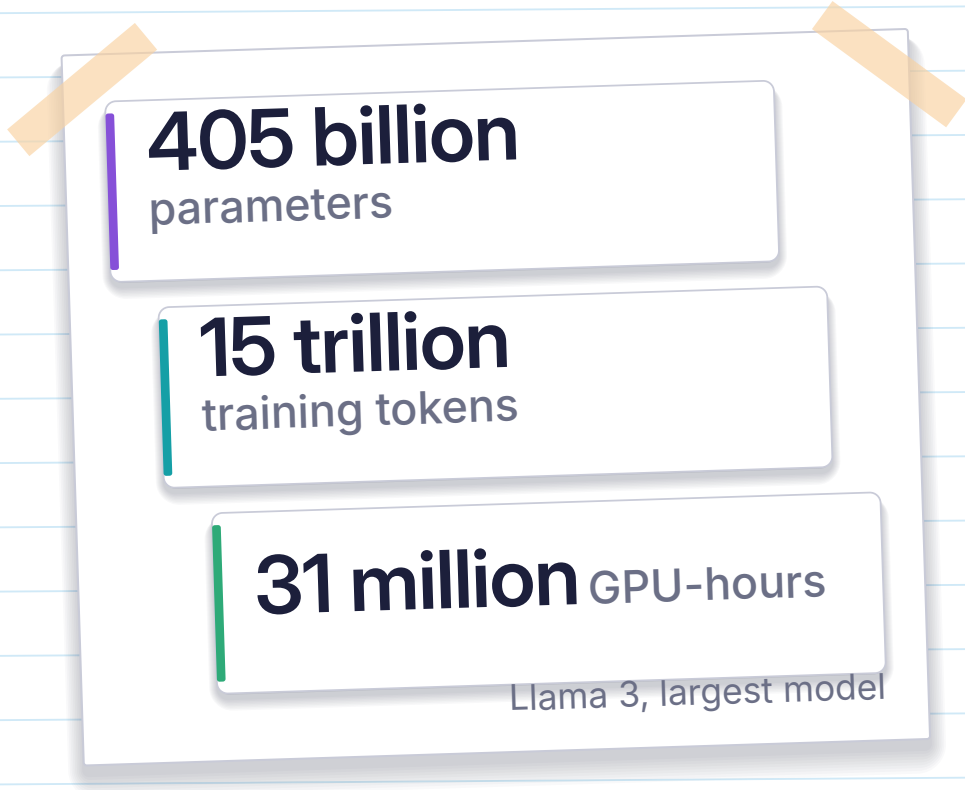
Prof. Aviv Zohar

Semester A, 2026–7

Training a Language Model at Scale

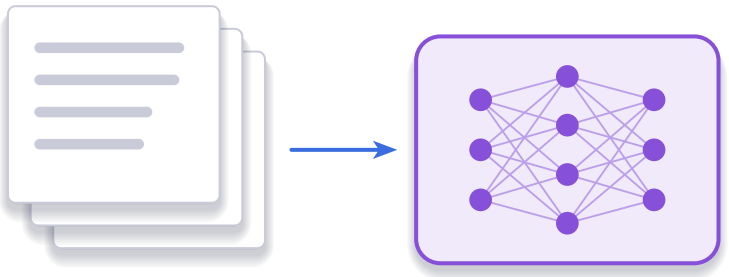
Post-training, size, data, compute, serving, and how we got here



A second training phase makes the assistant

① Pretraining

predict the next token



mostly public internet text

pretrained model

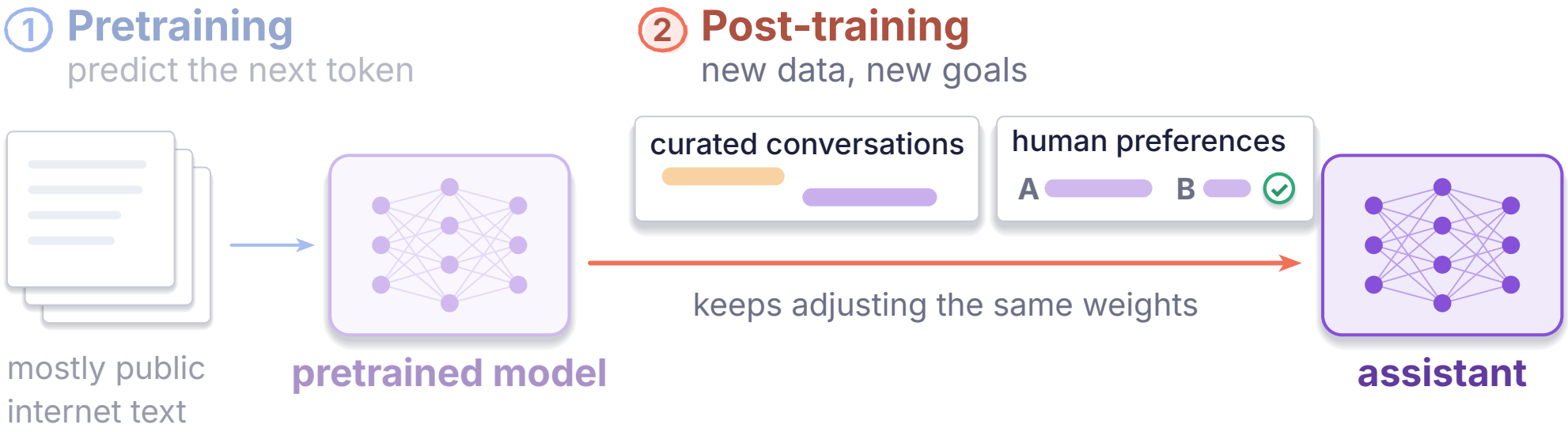
assistant?

prompt	`How do I reverse a list?`

continues	`How do I sort a list?`
	`How do I copy a list?`
	`How do I …`

More questions, not an answer.

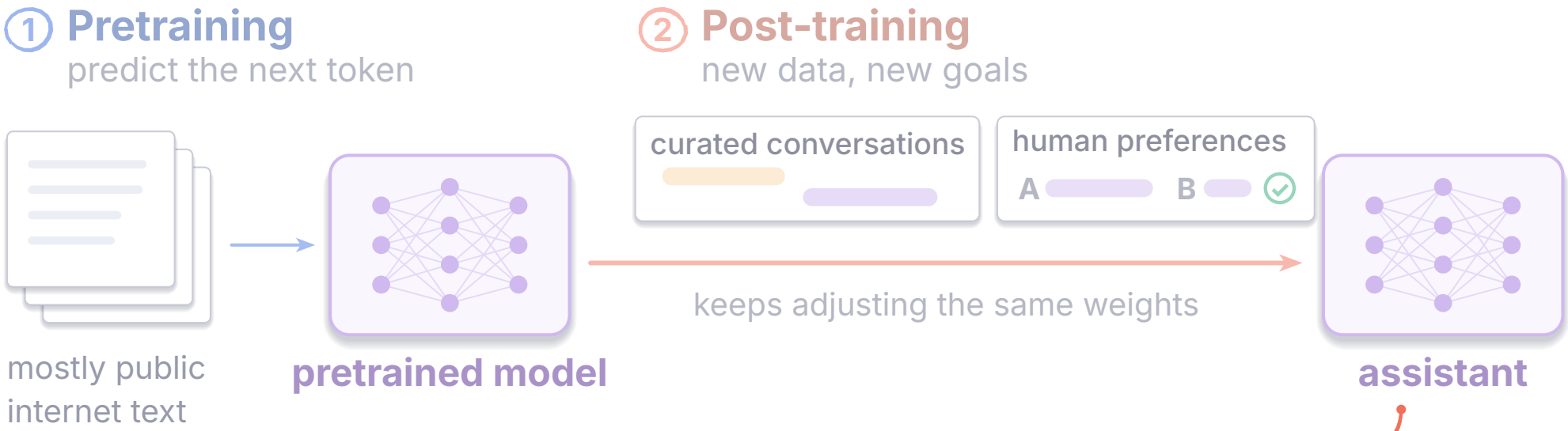
A second training phase makes the assistant



prompt	`How do I reverse a list?`

continues	`How do I sort a list?`
	`How do I copy a list?`
	`How do I …`

A second training phase makes the assistant



prompt	How do I reverse a list?
continues	How do I sort a list? How do I copy a list? How do I …

- helpful, polite replies
- follows instructions
- refuses harmful requests
- a consistent tone

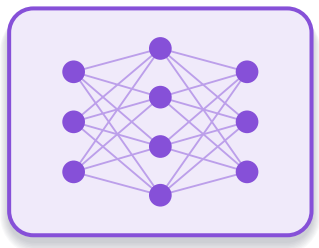
Encouraged, not guaranteed.

A second training phase makes the assistant

① Pretraining
predict the next token



mostly public internet text



pretrained model

knowledge, capabilities

```
prompt      How do I reverse a list?

continues   How do I sort a list?
            How do I copy a list?
            How do I …
```

② Post-training
new data, new goals

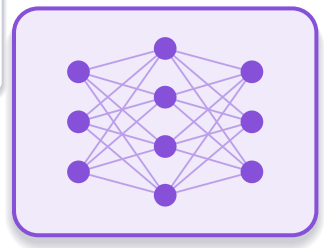
curated conversations

human preferences

A B

keeps adjusting the same weights

Both phases affect both.



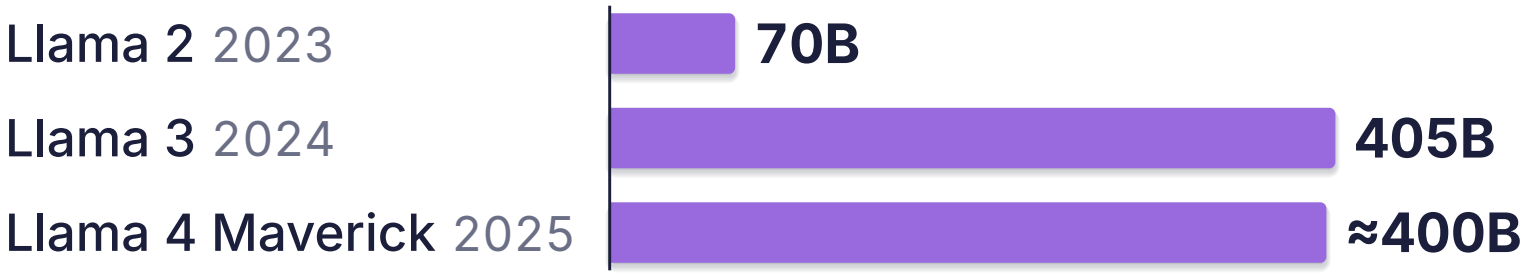
assistant

behavior as an assistant

- helpful, polite replies
- follows instructions
- refuses harmful requests
- a consistent tone

The model: parameters and vocabulary

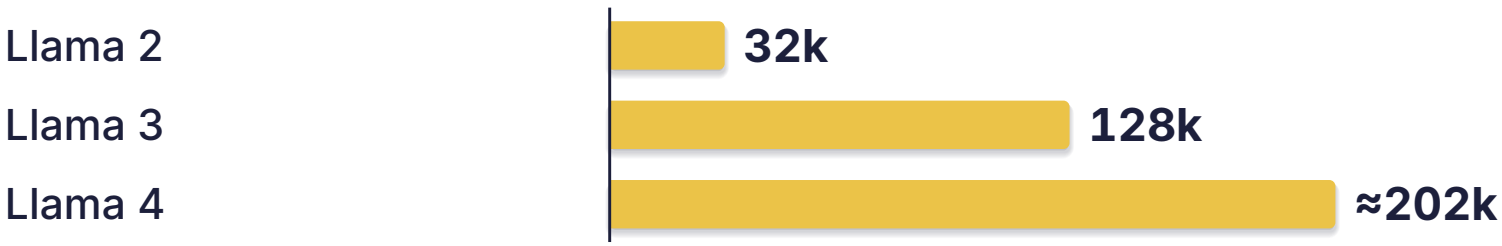
Parameters the largest model



human brain
≈100T
synapses

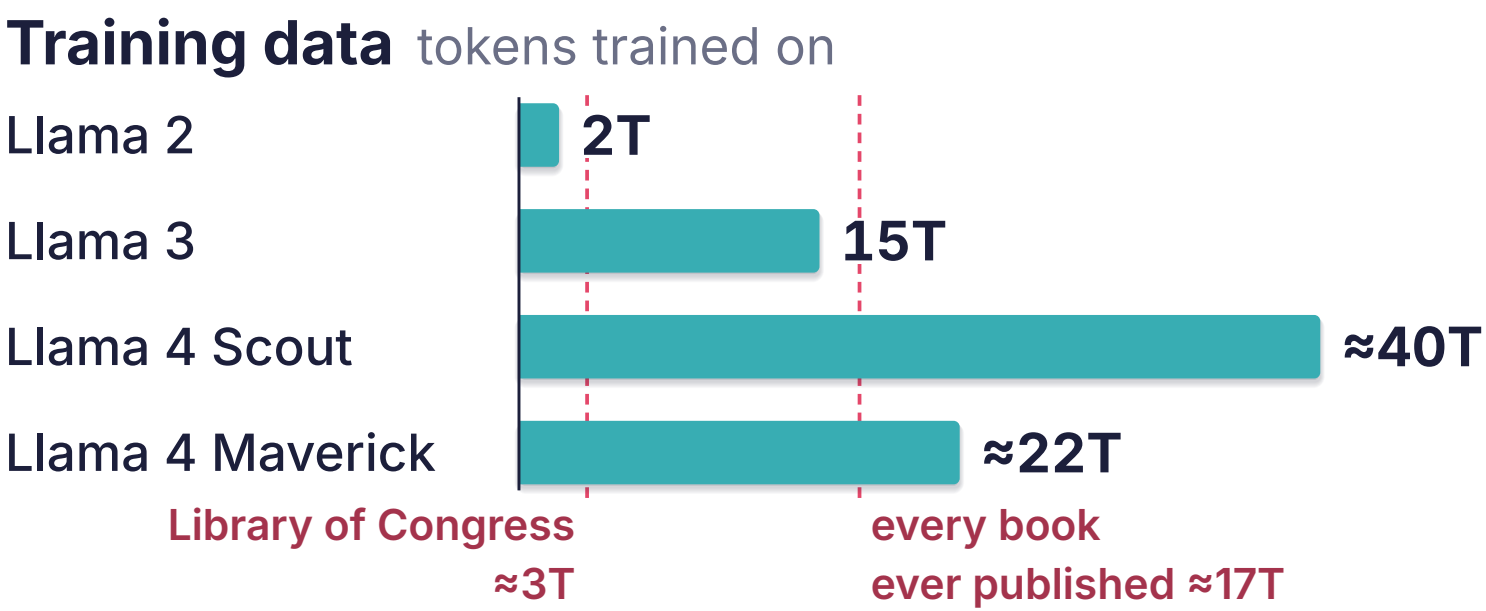
For scale, not a measure of intelligence.

Vocabulary tokens the model reads and writes



Meta: Llama 2, Llama 3 and Llama 4 releases

Training data: trillions of tokens



all public text online

≈300T

tokens, an estimate

In books

1 book ≈ 100,000 words
≈ 130,000 tokens

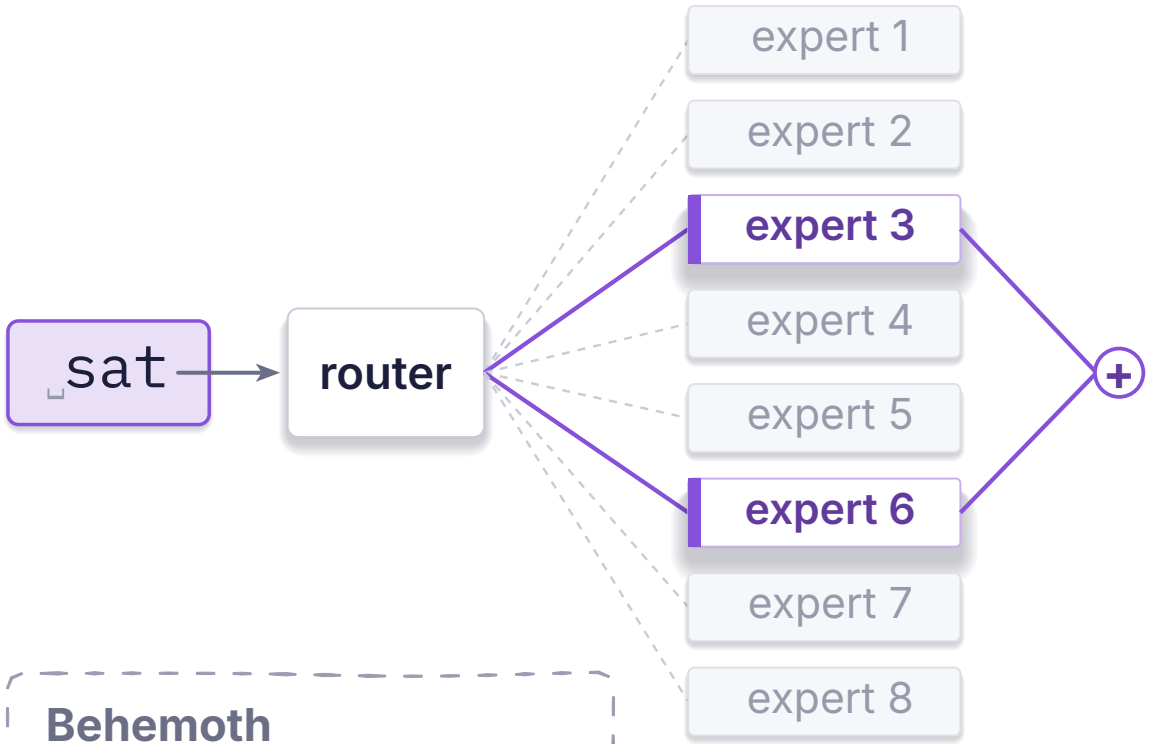
Library of Congress ≈ 26M books

every book ever published
≈ 130M (Google, 2010)

Scout and Maverick are two different Llama 4 models.

Meta; Library of Congress; Google Books; Epoch AI (2024)

A mixture of experts runs each token through some of its weights



Behemoth
nearly 2T parameters
still in training at release

Llama 4 Maverick
released April 2025, with Scout

stored: ≈400 billion parameters

active for each token: ≈17 billion

Less computation per token,
but all of it must be stored.

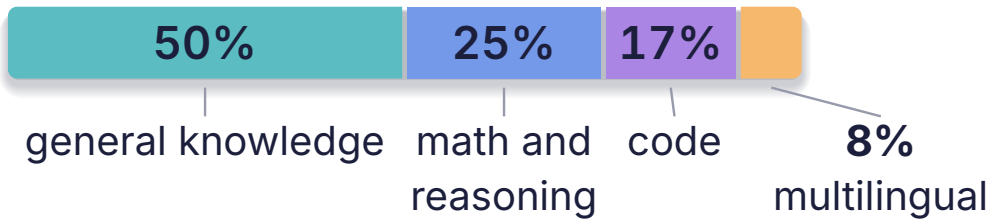
expert count and routing are illustrative

Where the training text comes from

Common sources

- **web crawls** — Common Crawl
- **code** — GitHub
- **Wikipedia**
- **books**
- **scientific papers**
- **Q&A sites**

Llama 3 · 15T tokens



Types of source, not every page or repository in one model.

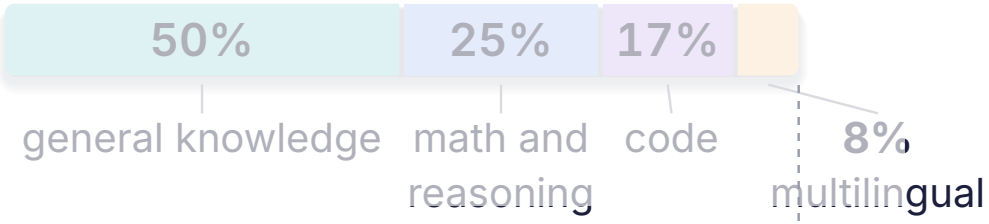
Meta: Llama 3 paper (2024), Llama 4 model card (2025)

Where the training text comes from

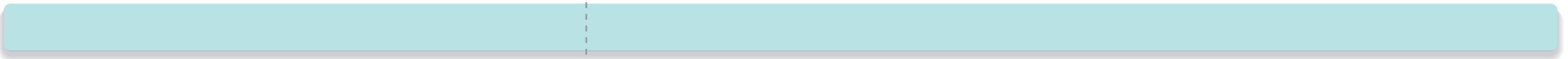
Common sources

- **web crawls** — Common Crawl
- **code** — GitHub
- **Wikipedia**
- **books**
- **scientific papers**
- **Q&A sites**

Llama 3 · 15T tokens



Llama 4 Scout · ≈40T tokens



- images and video
- ≈10× multilingual tokens · 200 languages
- public Instagram and Facebook posts
- interactions with Meta AI
- pretraining cutoff August 2024

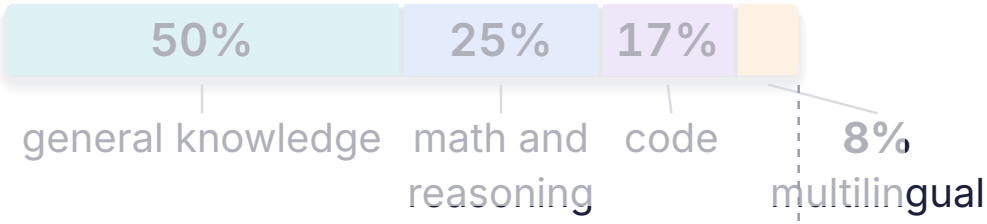
Meta: Llama 3 paper (2024), Llama 4 model card (2025)

Where the training text comes from

Common sources

- **web crawls** — Common Crawl
- **code** — GitHub
- **Wikipedia**
- **books**
- **scientific papers**
- **Q&A sites**

Llama 3 · 15T tokens



Llama 4 Scout · ≈40T tokens

no numerical breakdown published

≈25T more tokens: which sources, in what share?

images and video

≈10× multilingual tokens · 200 languages

public Instagram and Facebook posts

interactions with Meta AI

pretraining cutoff August 2024

Meta: Llama 3 paper (2024), Llama 4 model card (2025)

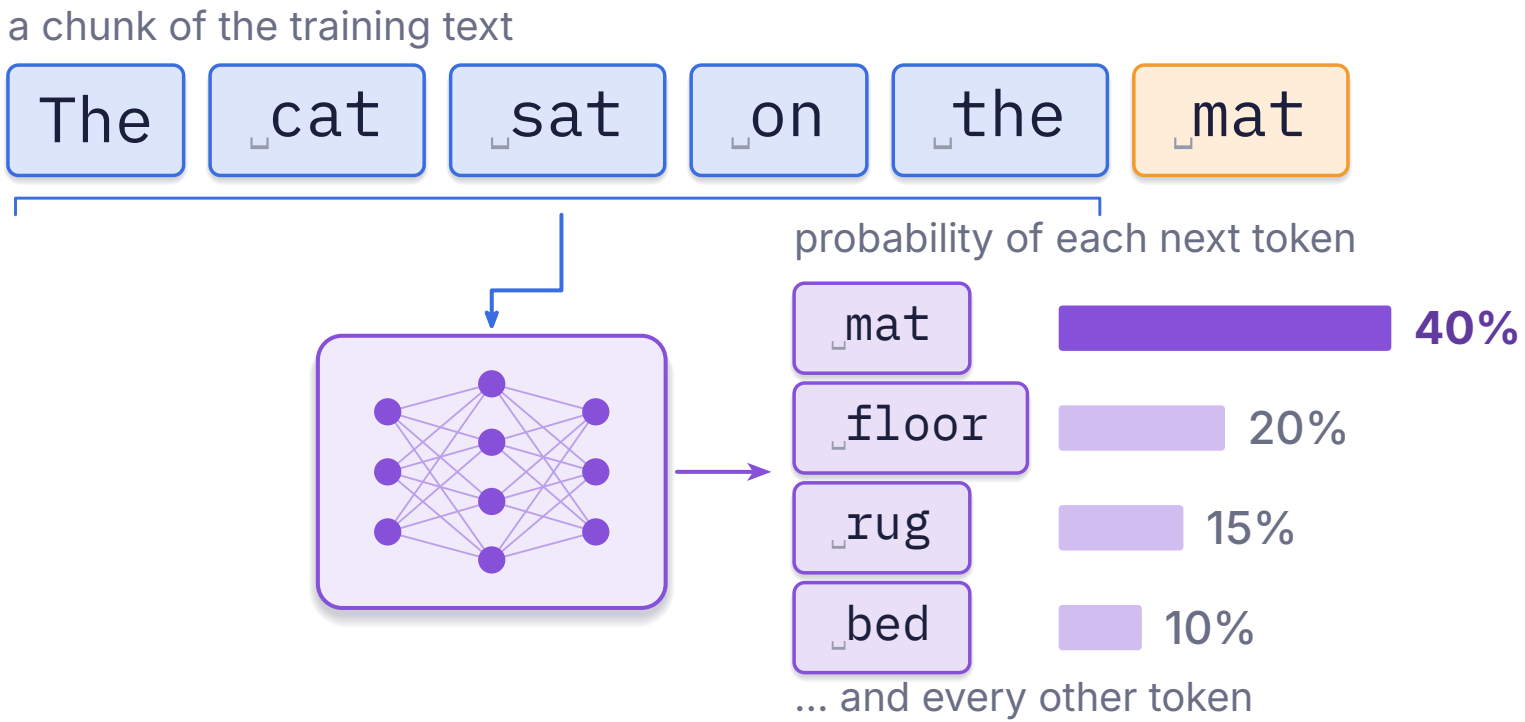
Every piece of text is a training example

a chunk of the training text

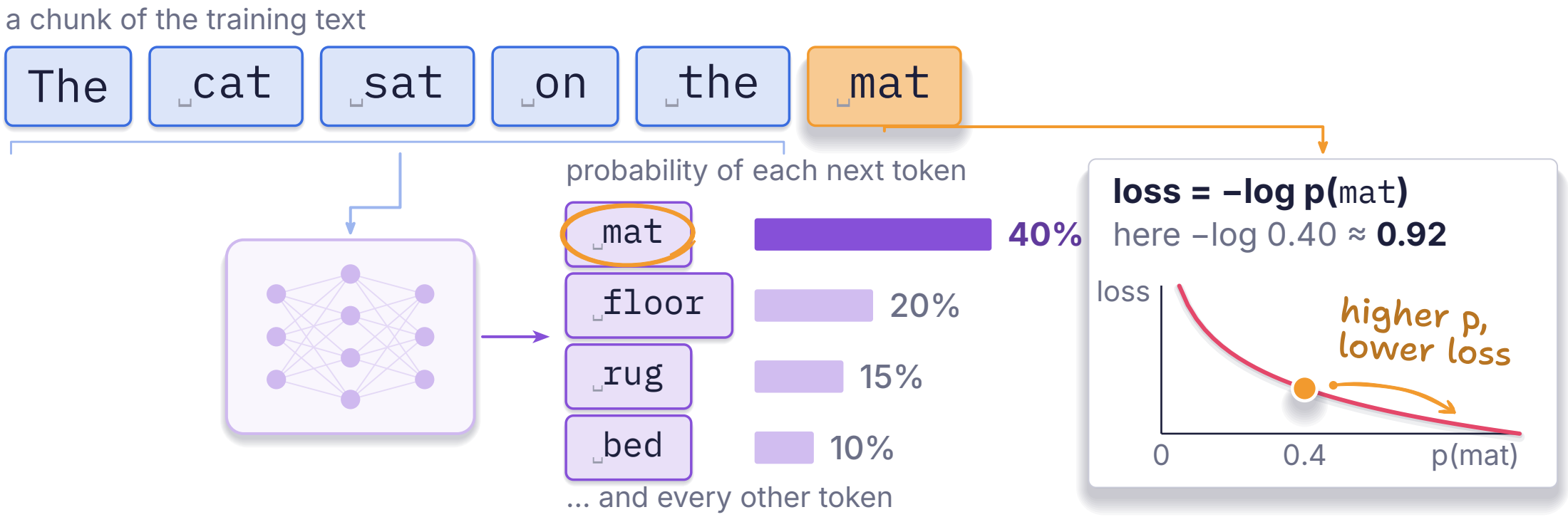


The correct next token, hidden from the model.

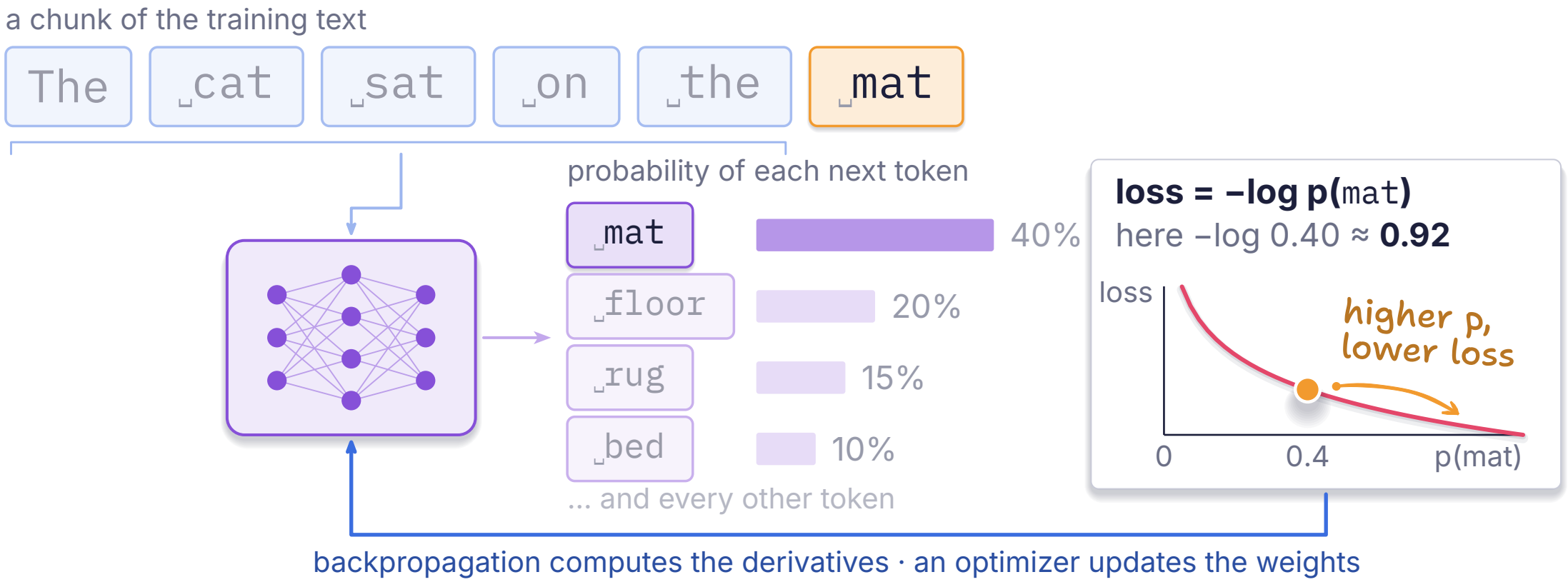
Every piece of text is a training example



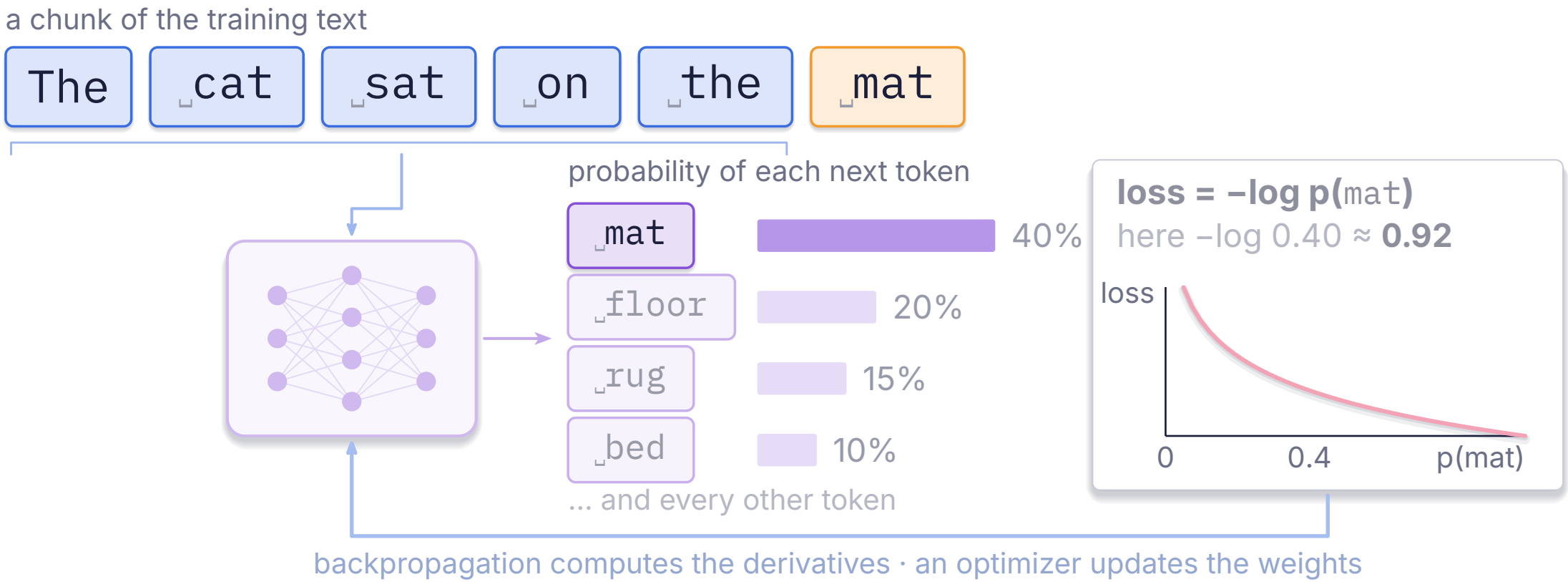
Every piece of text is a training example



Every piece of text is a training example



Every piece of text is a training example

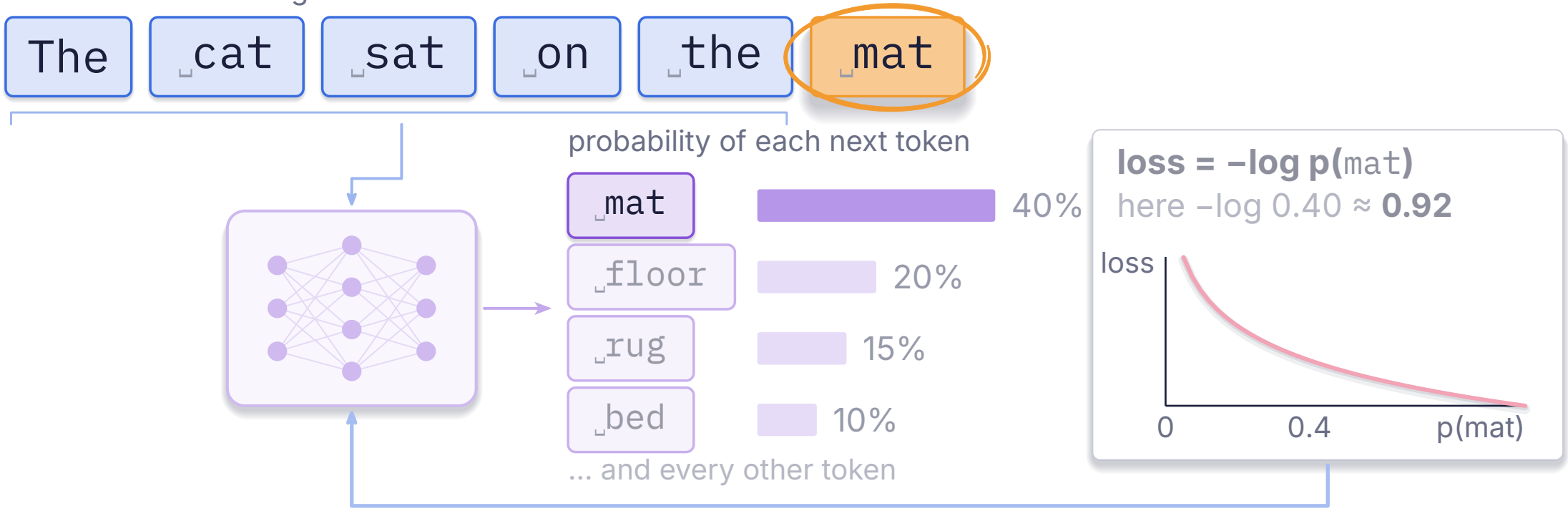


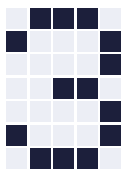
The cat sat on the mat

Repeated over billions of examples, many at a time: a batch.

Every piece of text is a training example

a chunk of the training text

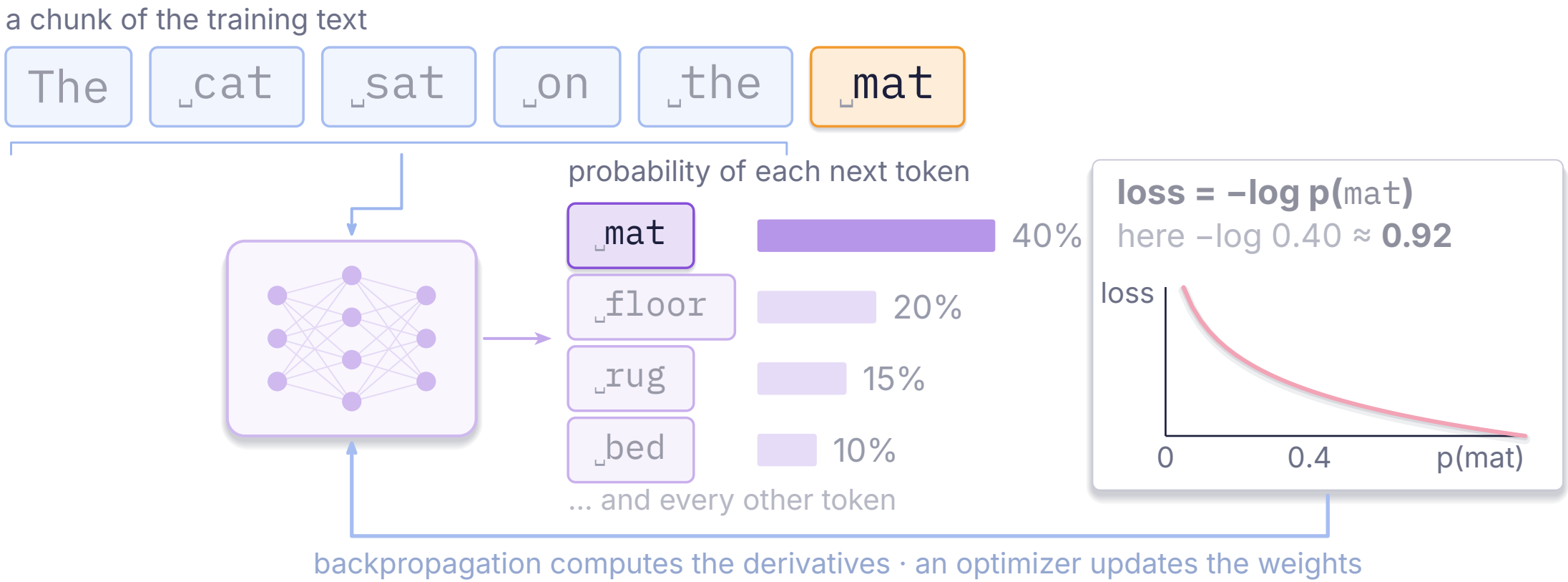




→ 3 someone supplies the label

The cat sat on the → ␣mat already in the text

Every piece of text is a training example



Self-supervised learning the training targets come from the data itself

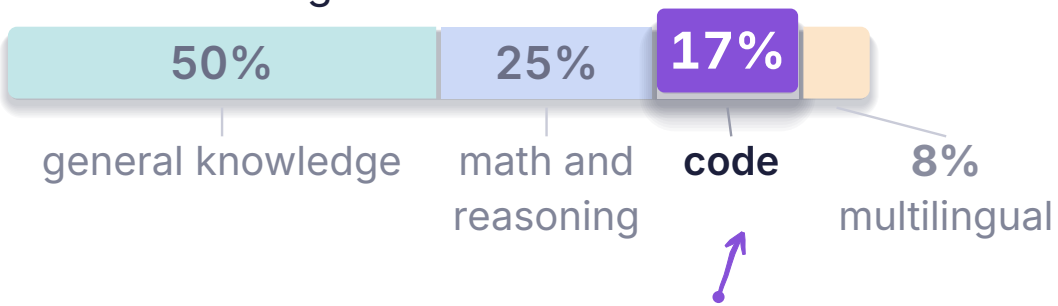
The data still costs money to gather, license, filter and process, and suitable fresh text is finite.

Coding suits the way these models are trained

① Plenty of examples

public repositories hold enormous amounts of code

Llama 3 training tokens

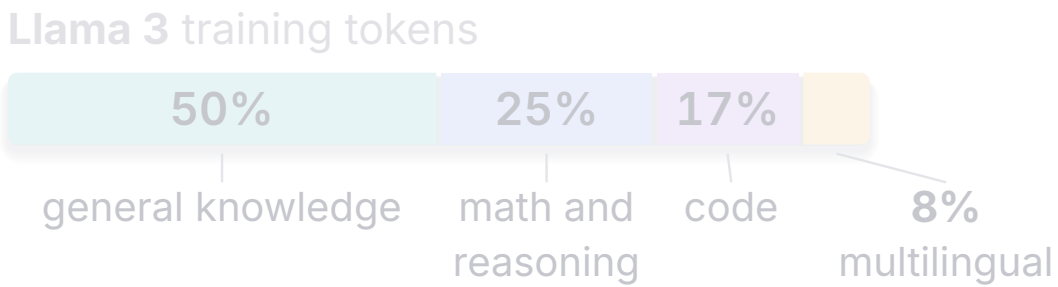


About 1 in 6 tokens was code.

Coding suits the way these models are trained

① Plenty of examples

public repositories hold enormous amounts of code

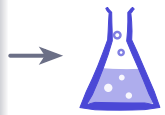


② A program can check the answer

run the solution against its tests

```
def reverse(xs):
    return xs[::-1]
```

a proposed solution



tests

reward

no reward

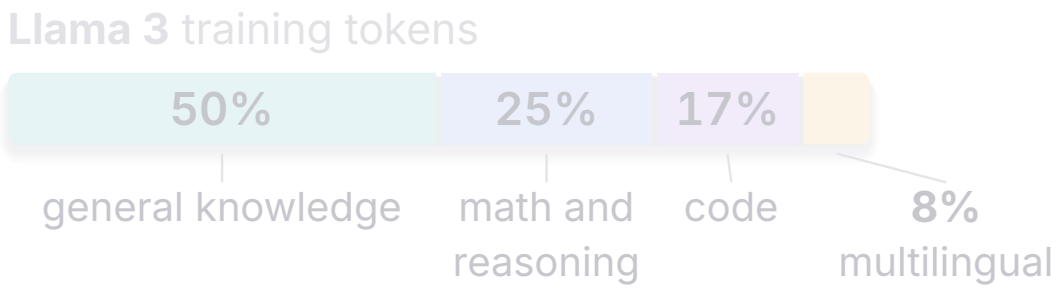
reinforcement learning with verifiable rewards

No human label for each example, so it can run on a huge number of problems.

Coding suits the way these models are trained

① Plenty of examples

public repositories hold enormous amounts of code



② A program can check the answer

run the solution against its tests

```
def reverse(xs):
    return xs[::-1]
```

a proposed solution



tests

reward

no reward

reinforcement learning with verifiable rewards

Expect gains wherever a program can check the answer.

Checkable answers help

- **an agent at work**
  runs code, reads the result, retries
- **math**
  answers checked automatically
- **proofs of theorems**
  checked by a proof checker

Training Llama 3 took about 31 million GPU-hours

H100 GPU-hours
≈31 million

GPUs, at most
≈16,000

days at that GPU count
≈80
two to three months

Llama 3, 405B parameters, as reported by Meta

Thought experiment: the same work on less hardware.
Order of magnitude only, not a benchmark.

Training Llama 3 took about 31 million GPU-hours

H100 GPU-hours
≈31 million

GPUs, at most
≈16,000

days at that GPU count
≈80 two to three months

Llama 3, 405B parameters, as reported by Meta

One GPU ≈3,500 years if it could hold the model in memory



Started 3,500 years ago, still running today.

Training Llama 3 took about 31 million GPU-hours

H100 GPU-hours	GPUs, at most	days at that GPU count
≈31 million	≈16,000	≈80 two to three months

Llama 3, 405B parameters, as reported by Meta

One GPU ≈3,500 years if it could hold the model in memory

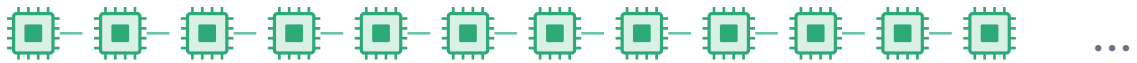


One CPU core ≈60 million years assuming it is 17,000 times slower than one GPU



A training run costs on the order of $100 million

Thousands of GPUs share the work and exchange results: months, not millennia.



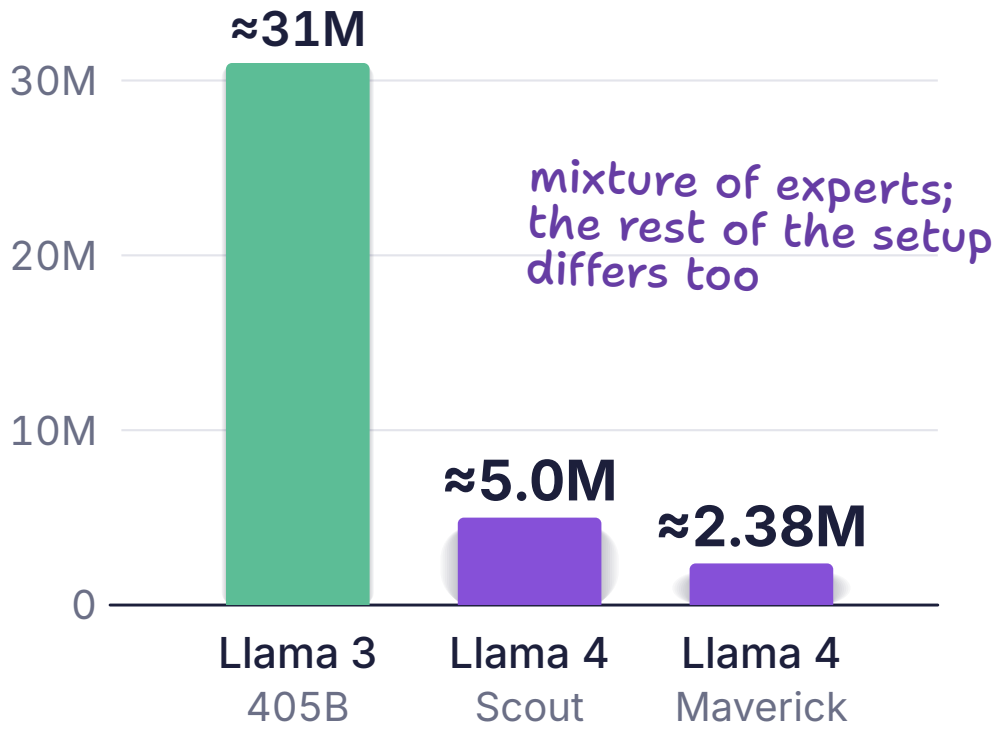
31 million GPU-hours at an assumed rental price

$3 per GPU-hour **≈$100 million**

$7 per GPU-hour **≈$220 million**

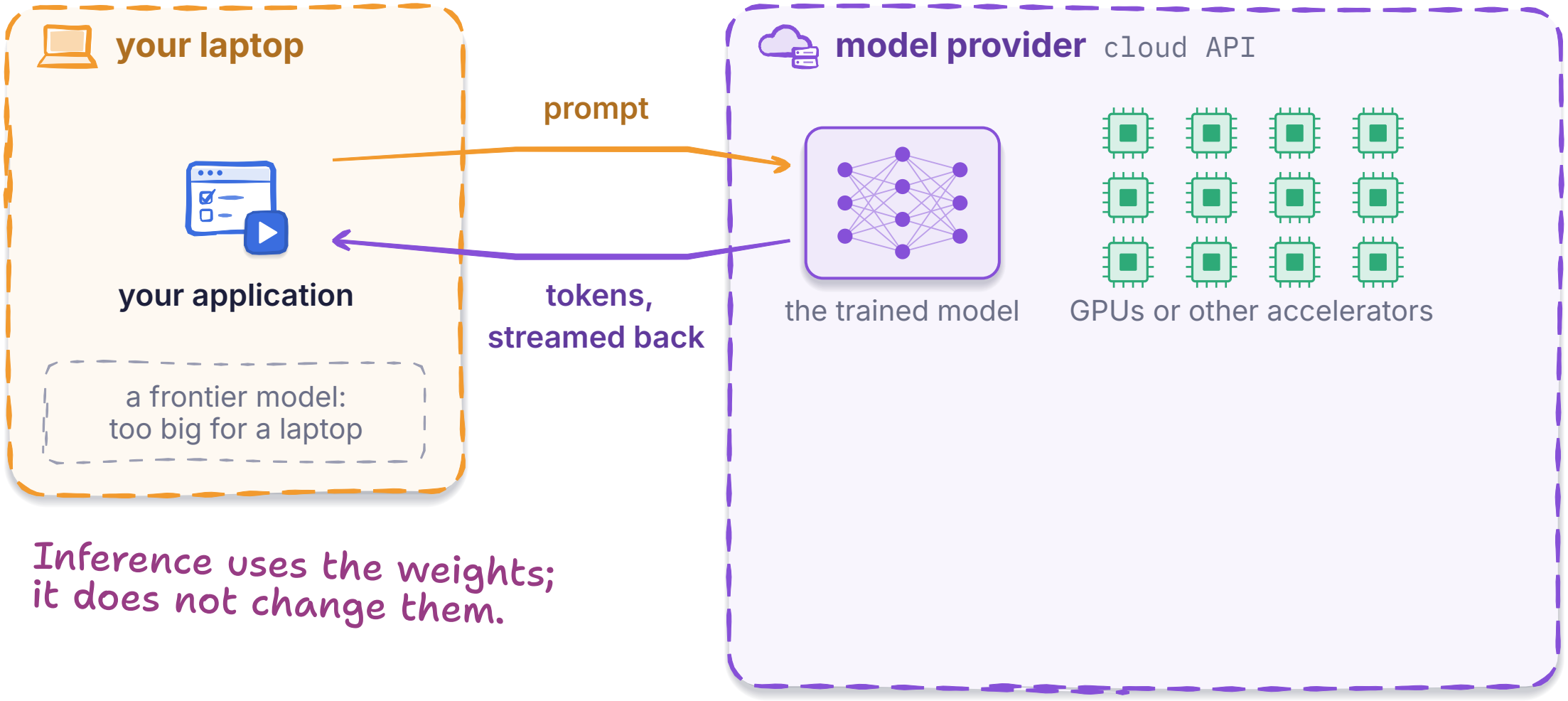
Assumed rates, not Meta's bill. GPU time for one run: experiments and other costs come on top.

GPU-hours per training run



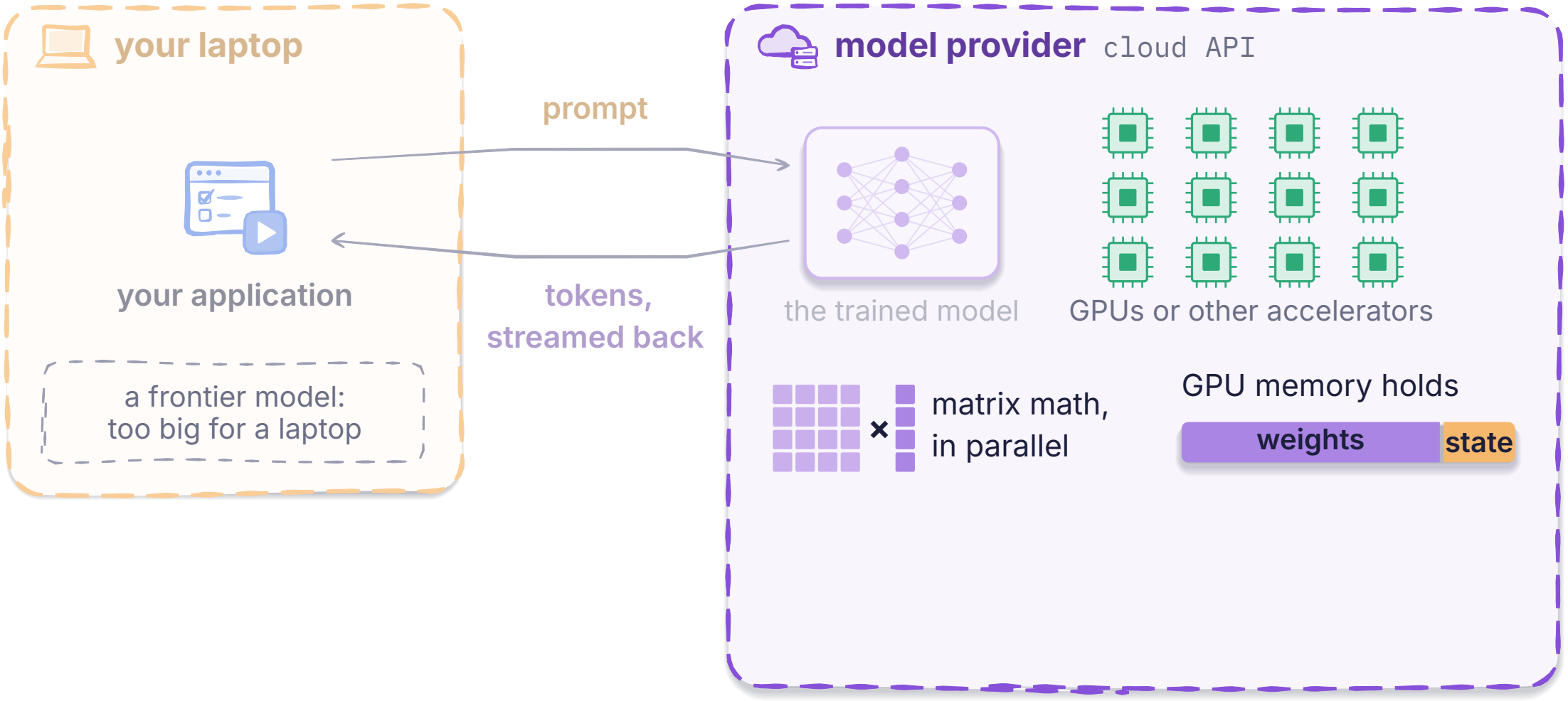
Meta model cards; rental rates assumed

A request travels to the model, not the model to you

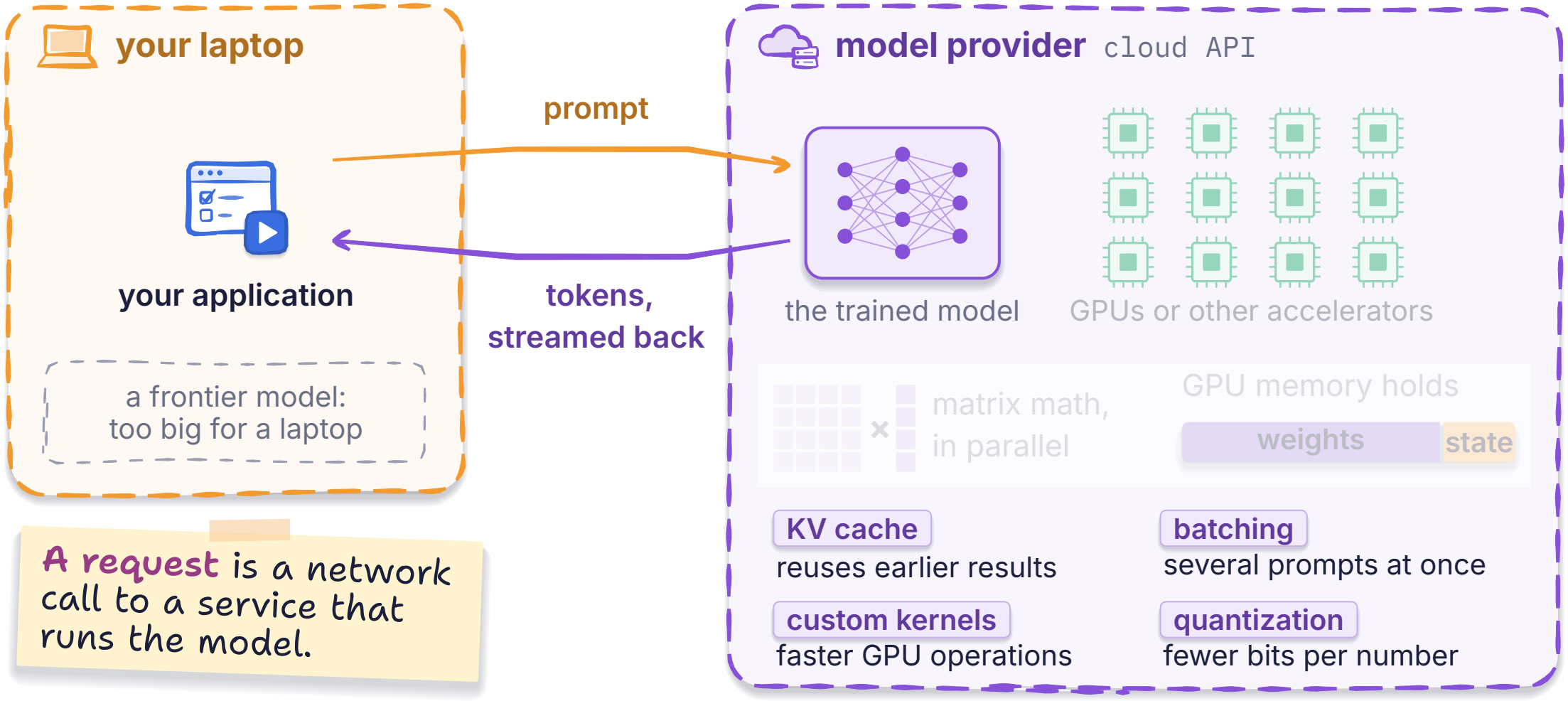


Inference uses the weights;
it does not change them.

A request travels to the model, not the model to you



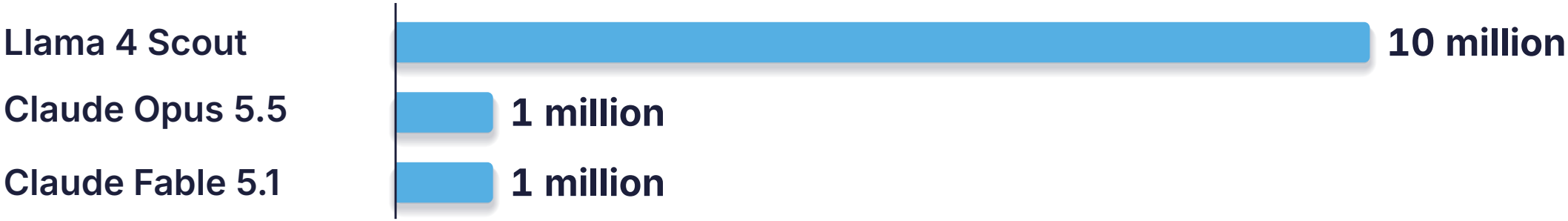
A request travels to the model, not the model to you



A request is a network call to a service that runs the model.

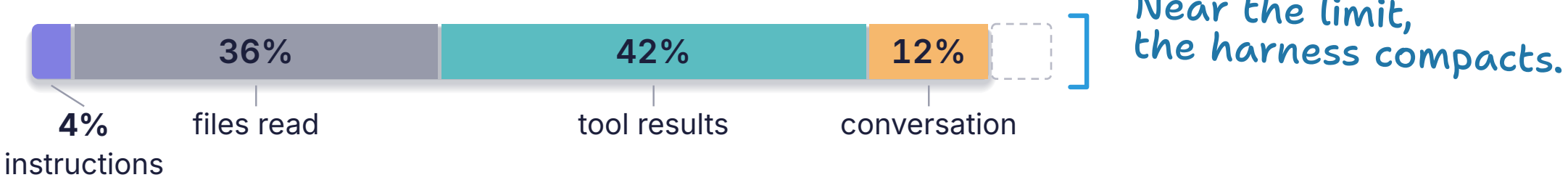
Context windows reach millions of tokens

The most tokens the model can take in one request



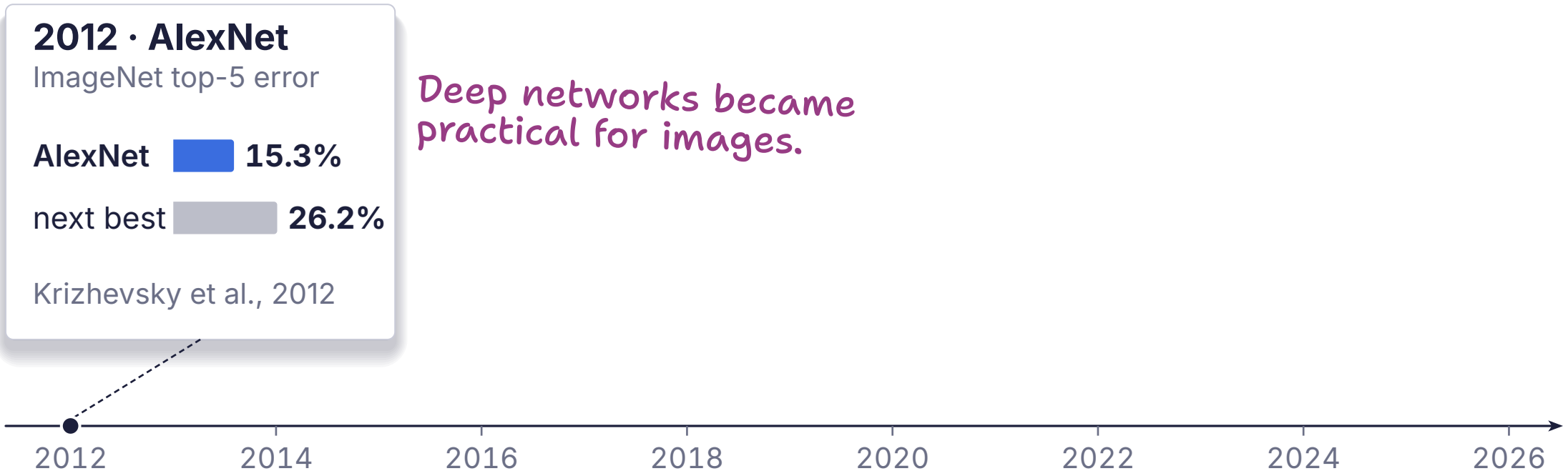
in the documented offerings; a given service may be configured lower

A long agent session fills the window

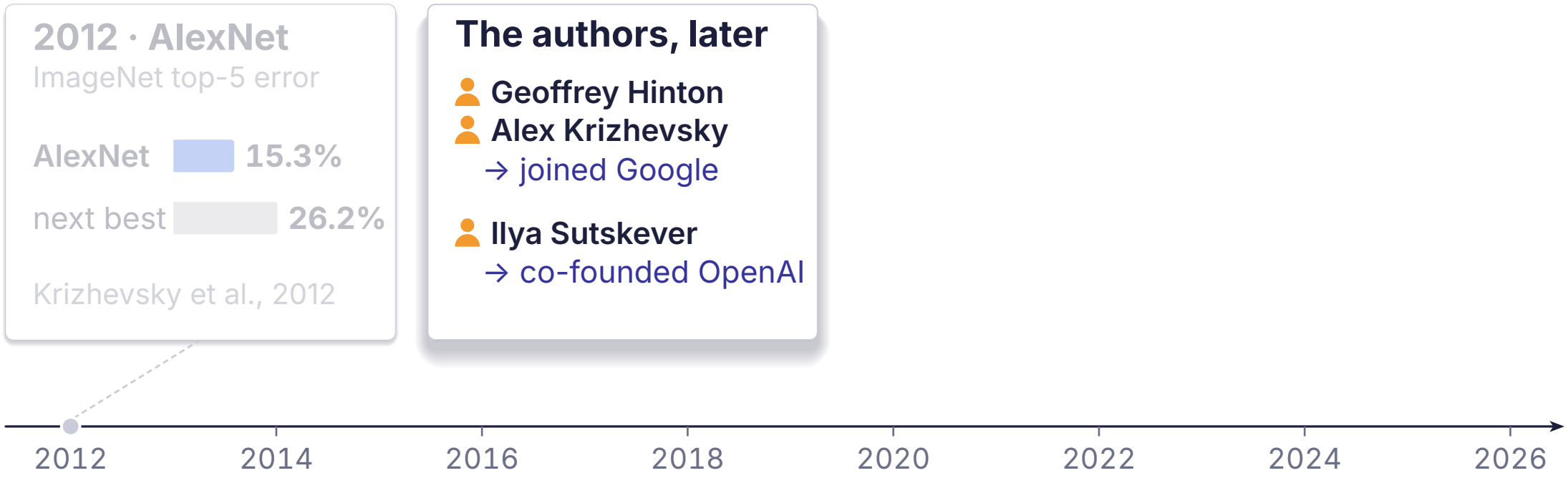


session shares are illustrative

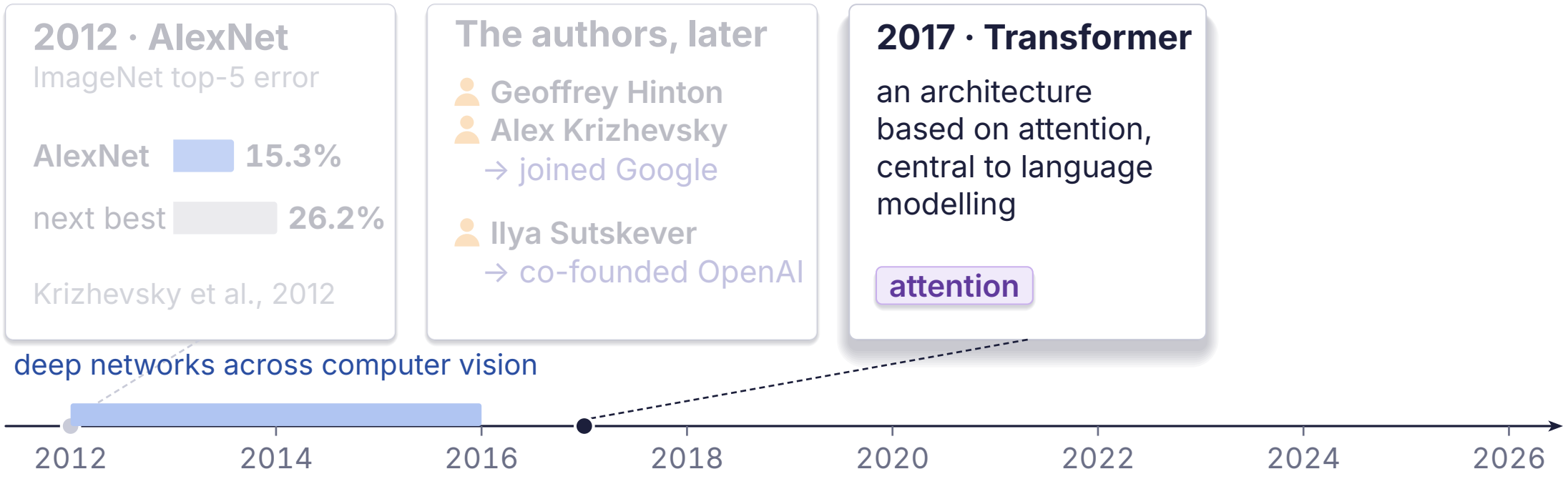
From AlexNet to coding agents



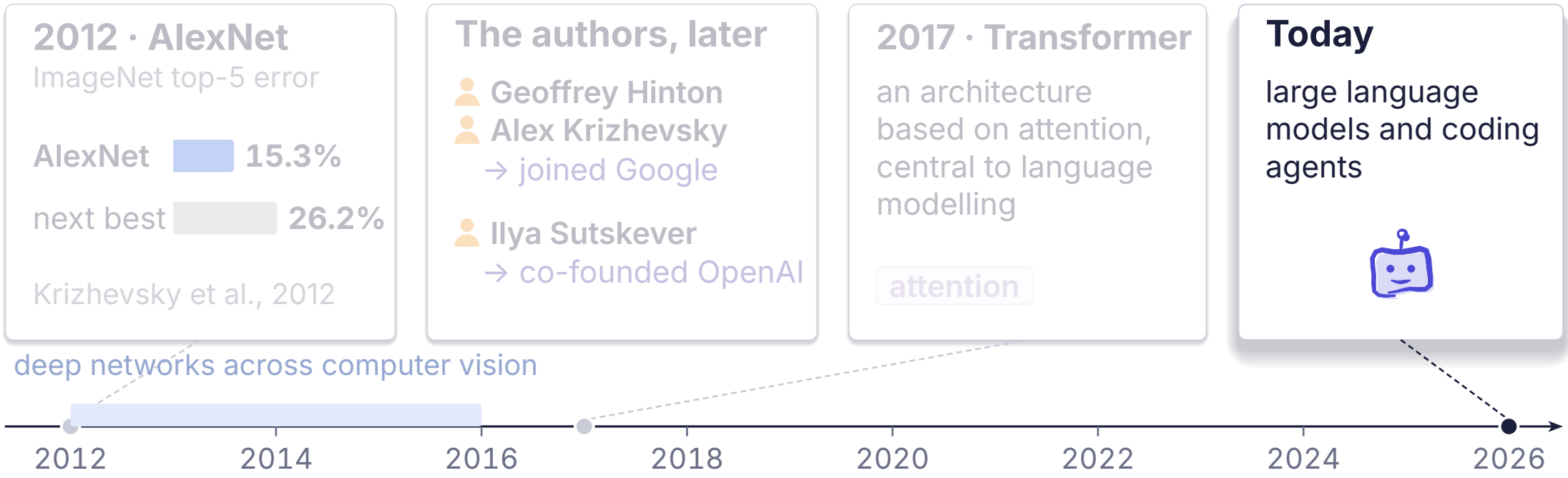
From AlexNet to coding agents



From AlexNet to coding agents



From AlexNet to coding agents



From AlexNet to coding agents

2012 · AlexNet

ImageNet top-5 error

AlexNet 15.3%

next best 26.2%

Krizhevsky et al., 2012

The authors, later

Geoffrey Hinton

Alex Krizhevsky

→ joined Google

Ilya Sutskever

→ co-founded OpenAI

2017 · Transformer

an architecture based on attention, central to language modelling

attention

Today

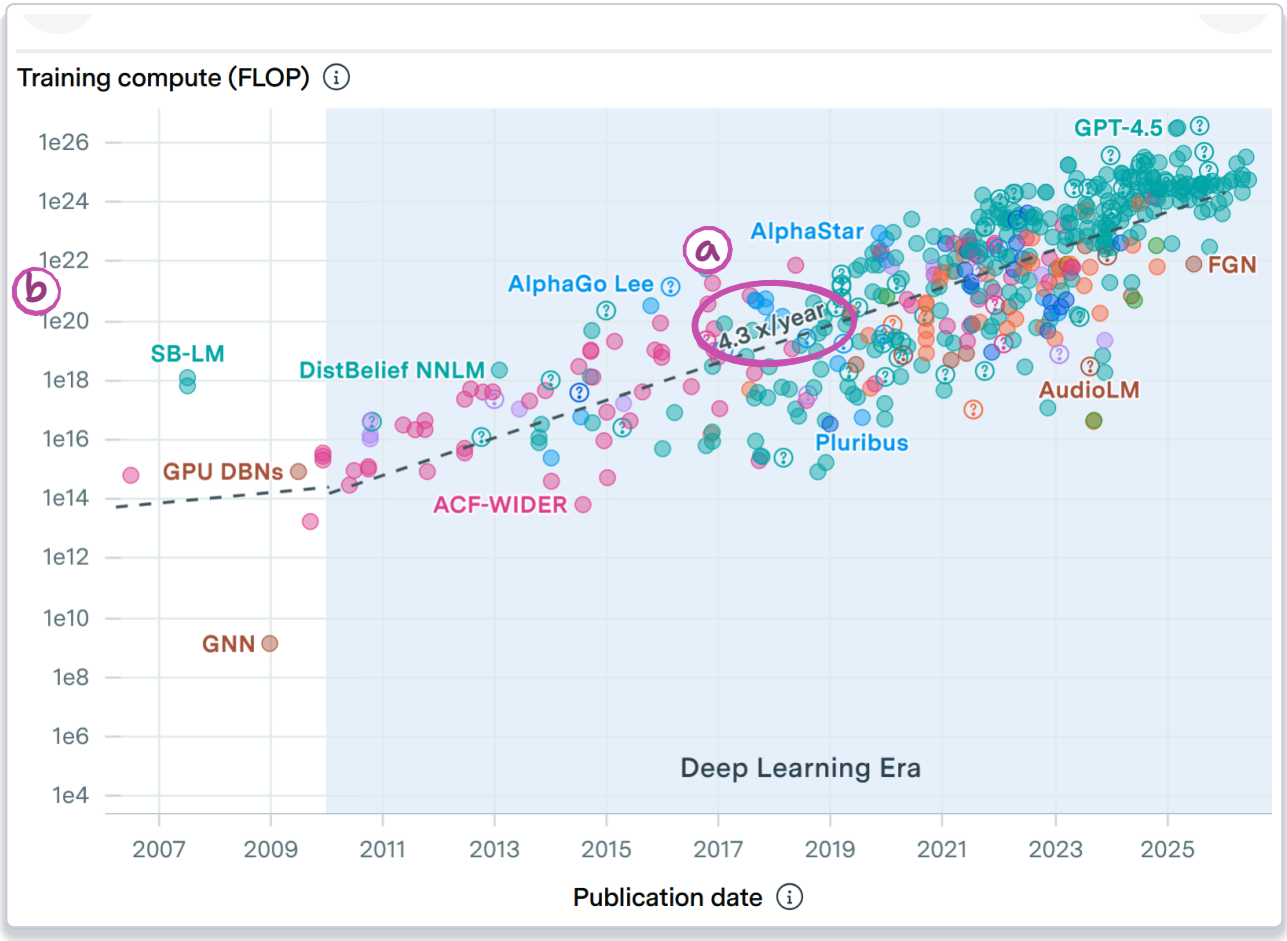
large language models and coding agents



more data and computation → larger networks, useful on harder tasks

Architecture and training methods improved alongside.

Training computation has grown very fast



screenshot · Epoch AI, Data on Notable AI Models

growth per year

4–5×

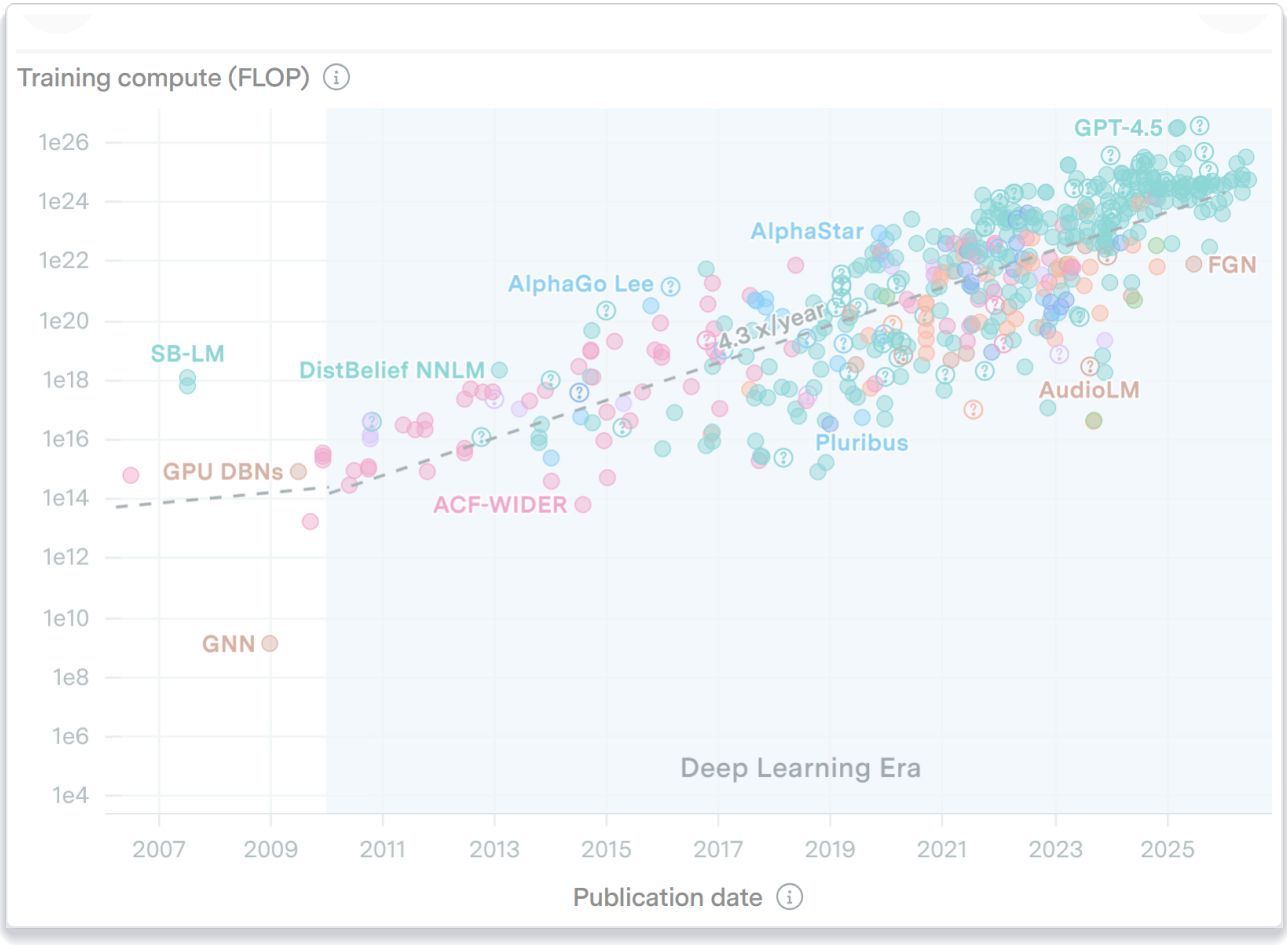
2010 to 2024 · Epoch AI, 2024

ⓐ each point is a notable model

ⓑ log scale: 1e22 is 100 × 1e20

fitted trend → rapid growth

Training computation has grown very fast



screenshot · Epoch AI, Data on Notable AI Models

growth per year

4–5×

2010 to 2024 · Epoch AI, 2024

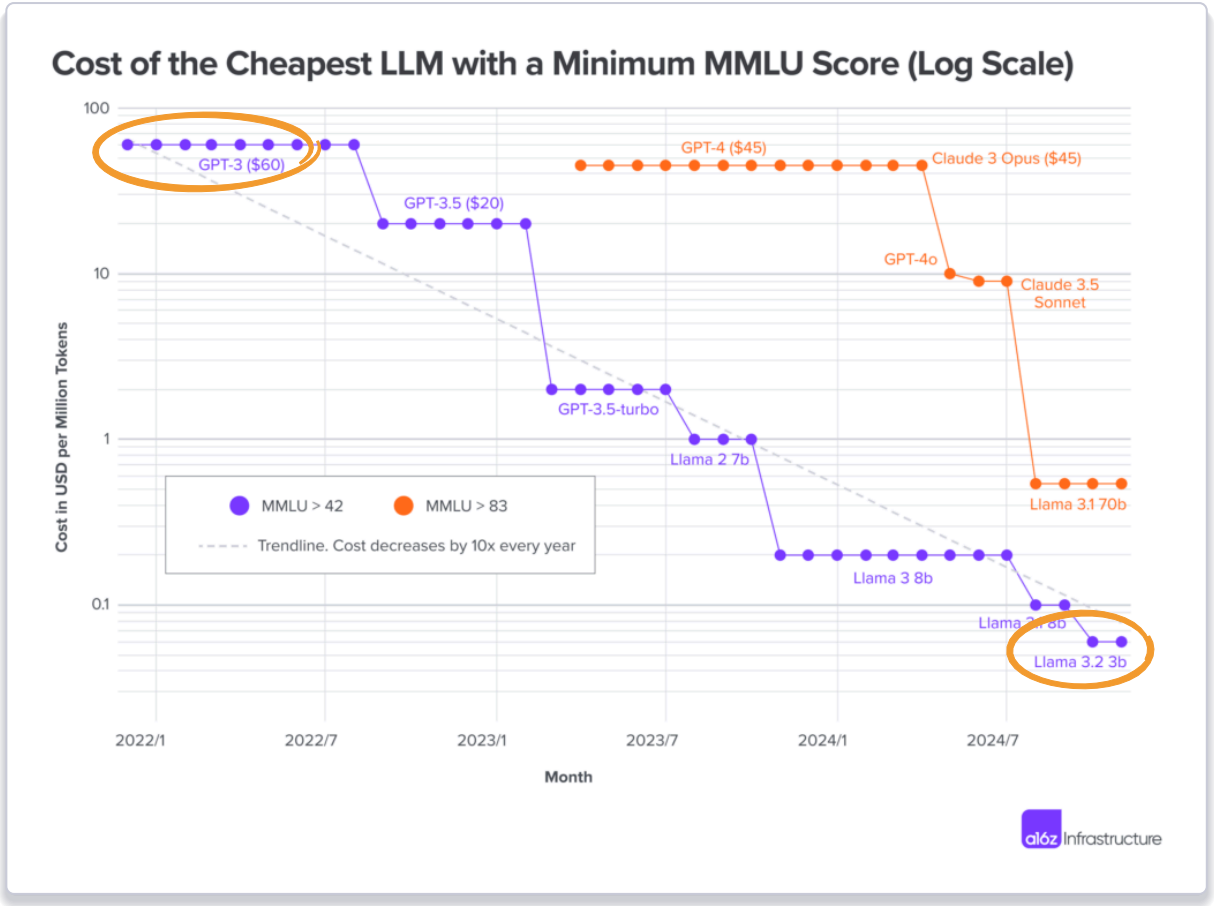
since 2010

≈10,000,000,000×

about ten orders of magnitude · Epoch AI

Much faster than single chips improved: labs used more chips, time and money.

A fixed level of performance keeps getting cheaper



screenshot · a16z, "Welcome to LLMflation" (2024)

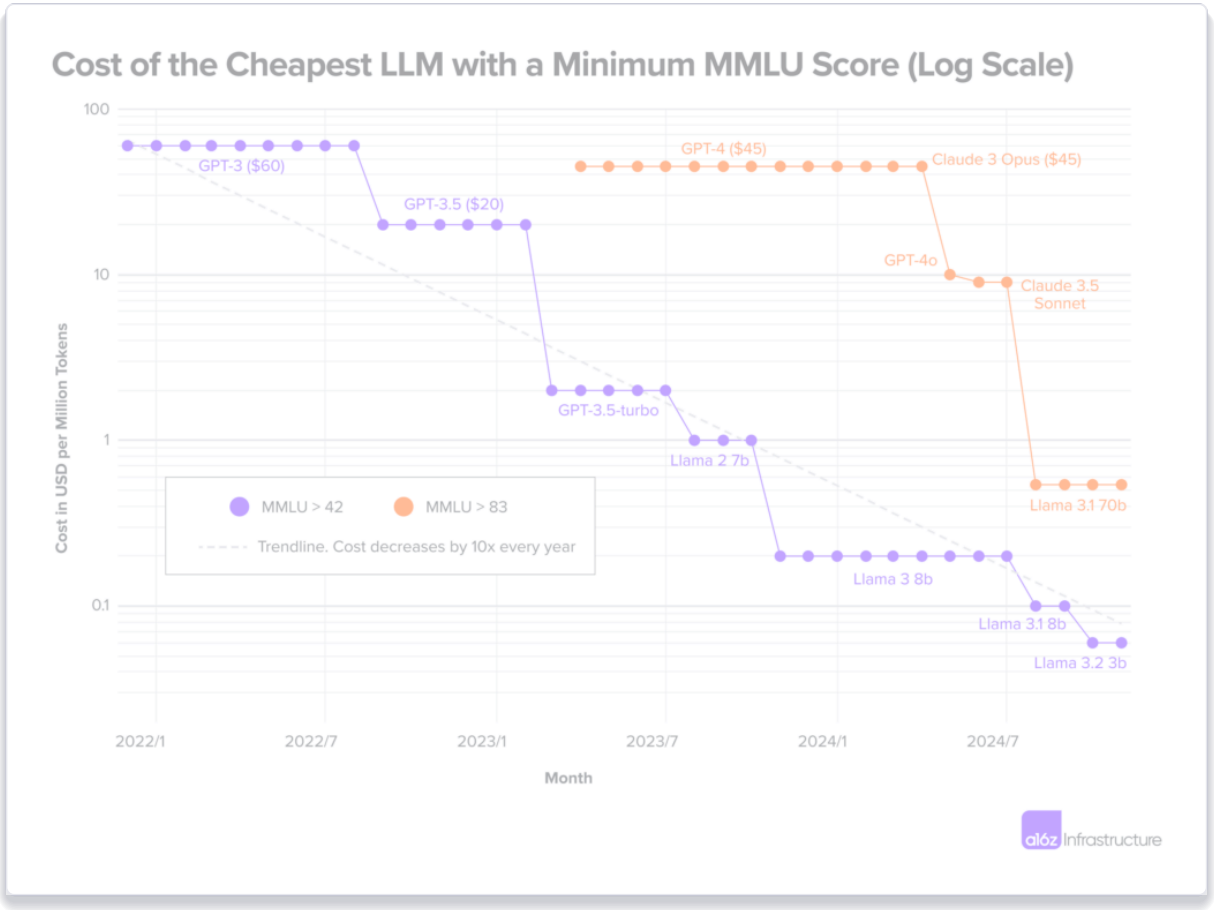
Fixed score: MMLU 42
the cheapest model that reaches it

		per million tokens
GPT-3	Nov 2021	**$60**
Llama 3.2 3B	late 2024	**$0.06**

cheaper for the same score
≈1000× in about 3 years

The price axis is logarithmic:
each step down is 10× cheaper.

A fixed level of performance keeps getting cheaper



screenshot · a16z, “Welcome to LLMflation” (2024)

Why a fixed level got cheaper

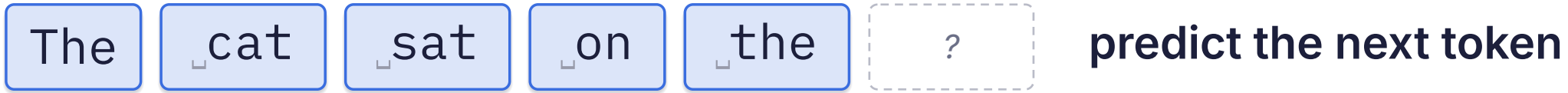
Better training methods
smaller models reach older scores

Cheaper serving
caching, batching, quantization

Competition
between providers lowers prices

An agent may call a model hundreds of times for one task. The bill still depends on the model, the context and the tokens generated.

A simple objective at a very large scale



predict the next token

A simple objective at a very large scale



parameters
405 billion

training tokens
≈15 trillion

GPU-hours
≈31 million

Llama 3, largest model · the text supplies its own targets, so no human labels each prediction

A simple objective at a very large scale



predict the next token

parameters
405 billion

training tokens
≈15 trillion

GPU-hours
≈31 million

Llama 3, largest model · the text supplies its own targets, so no human labels each prediction

Pretraining

broad capabilities

→

Post-training

an assistant's behavior

→

Inference

available to our programs