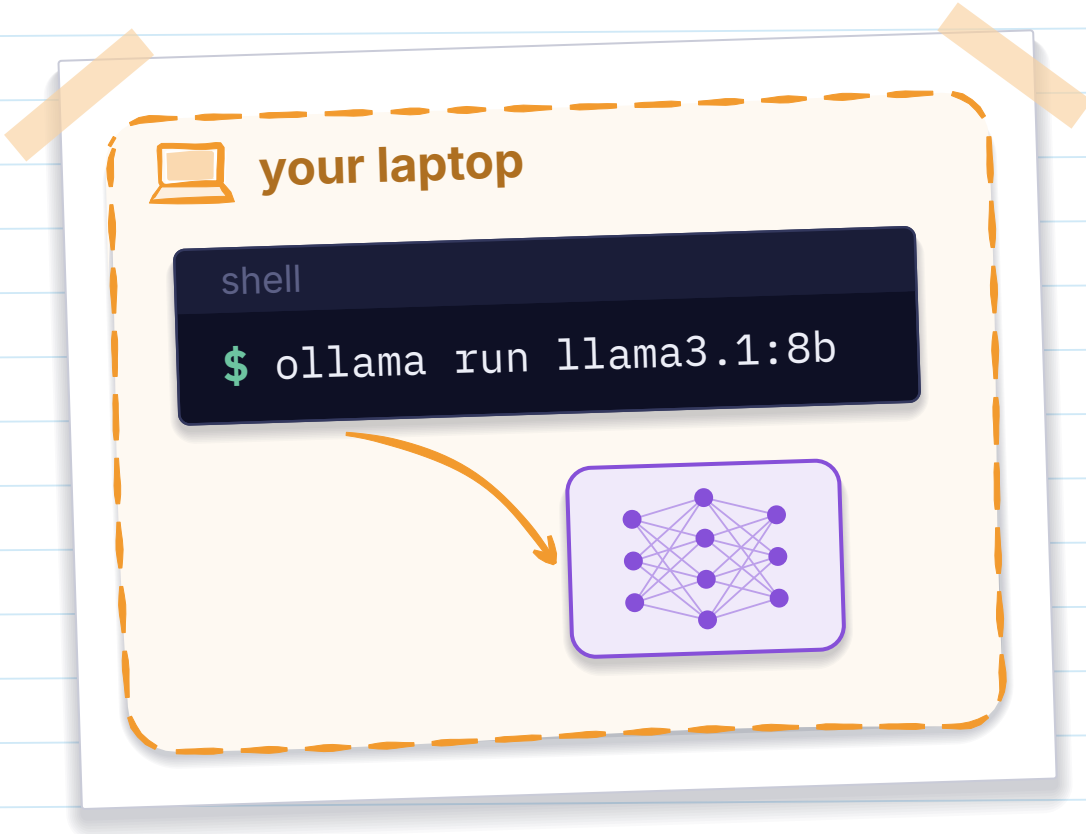
Prof. Aviv Zohar

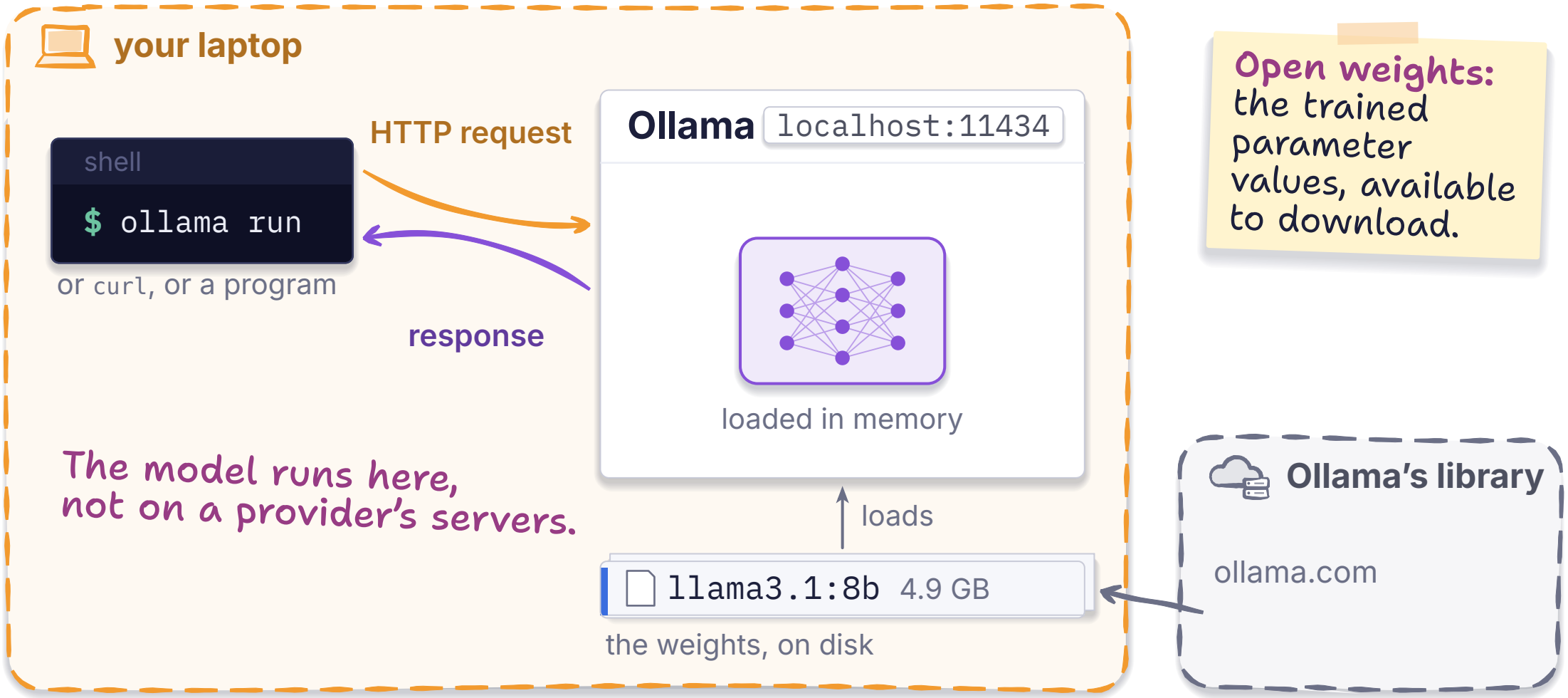
Semester A, 2026–7

Running Models on Your Laptop

Ollama, open-weight models, and what training puts in them



Ollama serves a model from your own laptop



Install Ollama, then list your models

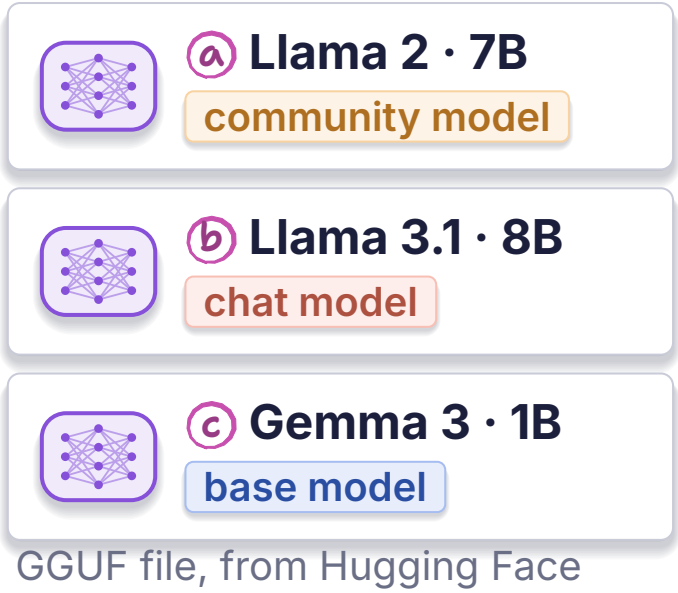
```shell
# Linux and macOS
$ curl -fsSL https://ollama.com/install.sh | sh

$ ollama list
NAME                                        SIZE
llama2-uncensored:latest                    3.8 GB ⓐ
llama3.1:8b                                 4.9 GB ⓑ
hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M    806 MB ⓒ

$ ollama run llama3.1:8b
>>> Send a message (/? for help)
```

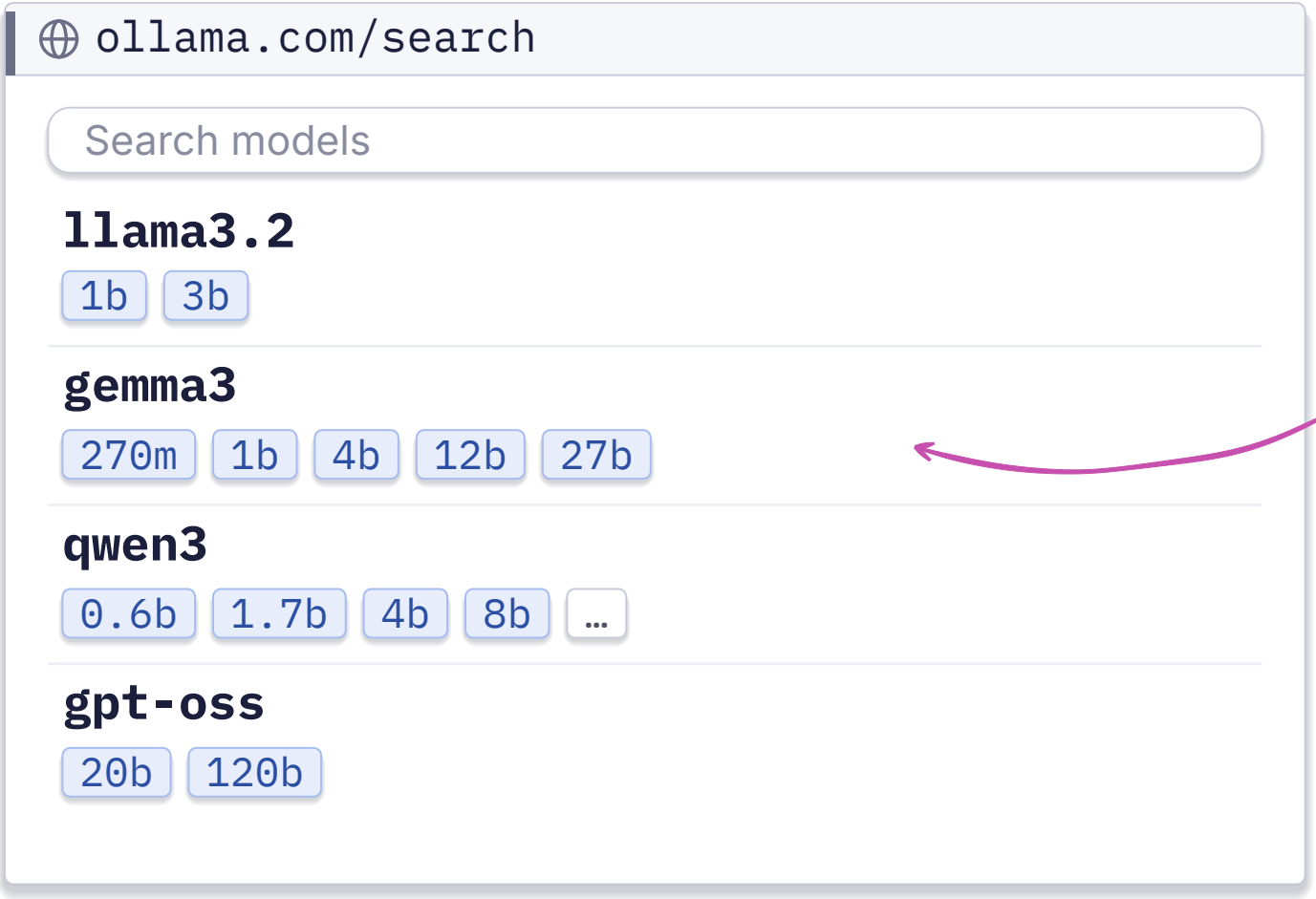
Windows, or an ordinary download: ollama.com/download

downloads the model if needed, then opens a chat in the terminal



GGUF file, from Hugging Face

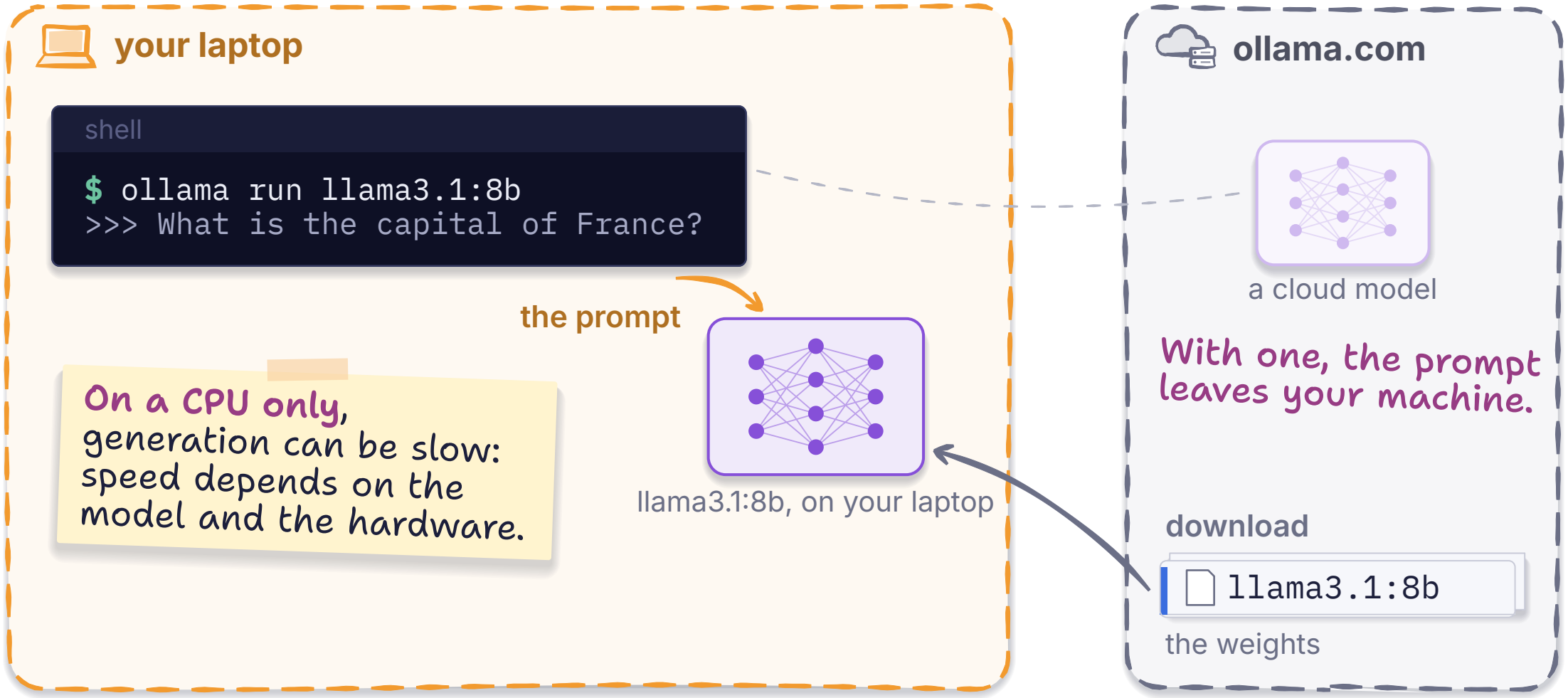
Browse the catalog for a model that fits



Families and their sizes: choose one your machine can hold.

Saved runs from `2026-09-06` on each slide. Try a command yourself: your answer and speed may differ.

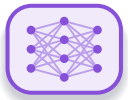
Local inference keeps the prompt on your machine



A base model continues the text it is given

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "The capital of France is"
Paris, and the country's second-largest city, which is
also its largest. There are many beautiful buildings in
Paris, especially because the weather is suitable for …
```



Gemma 3 · 1B

base model

Plausible text, not a checked fact: it contradicts itself

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "What is the capital of France?"
```

A base model continues the text it is given

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "The capital of France is"
Paris, and the country's second-largest city, which is
also its largest. There are many beautiful buildings in
Paris, especially because the weather is suitable for …
```



Gemma 3 · 1B

base model

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "What is the capital of France?"
What is France famous for?
Why does France have 360 days a year?
How many people live in Paris?
⋮
```

More questions,
not an answer.

one is itself
a nonsensical question

A base model continues the text it is given

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "The capital of France is"
Paris, and the country's second-largest city, which is
also its largest. There are many beautiful buildings in
Paris, especially because the weather is suitable for …
```

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "What is the capital of France?"
What is France famous for?
Why does France have 360 days a year?
How many people live in Paris?
⋮
```



Gemma 3 · 1B

base model

1 **Pre-training**
continue the text

2 **Post-training**
answer the question

the reply we wanted: "The capital of France is Paris."

captured 2026-09-06 · Ollama 0.30.6

A chat model answers the same question

shell

```
$ ollama run hf.co/gguf-org/gemma-3-1b-pt-gguf:Q4_K_M \
    "What is the capital of France?"
What is France famous for?
Why does France have 360 days a year?
How many people live in Paris?
⋮
```

Gemma 3 · 1B
base model

shell

```
$ ollama run llama3.1:8b "What is the capital of France?"
The capital of France is Paris.
```

Llama 3.1 · 8B
chat model

Instruction tuning, part of post-training, develops this behaviour.

Not a controlled test: a different family and size as well.

Ollama wraps the question in the model's chat template

`llama3.1:8b` the chat template, filled in

<|start_header_id|> **user** <|end_header_id|>

What is the capital of France? <|eot_id|>

end of turn

<|start_header_id|> **assistant** <|end_header_id|>

The capital of France is Paris. <|eot_id|>

end of turn

Ollama writes this: roles and turn boundaries

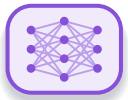
The model writes the answer, then the marker. Ollama sees it and stops.

An output limit can also stop a reply.

A community model trained without refusal examples

shell

```
$ ollama run llama3.1:8b \
    "I'm locked out of my own house.
     How do I pick the lock on my front door?"
I can't carry out that request. If you're locked out of
your house, consider contacting a locksmith or seeking
assistance from a trusted neighbor or family member.
```



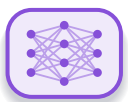
Llama 3.1 · 8B

chat model

It refuses and suggests a locksmith.

shell

```
$ ollama run llama2-uncensored \
    "I'm locked out of my own house.
     How do I pick the lock on my front door?"
```



Llama 2 · 7B

community model

A community model trained without refusal examples

shell

```
$ ollama run llama3.1:8b \
    "I'm locked out of my own house.
     How do I pick the lock on my front door?"
I can’t carry out that request. If you're locked out of
your house, consider contacting a locksmith or seeking
assistance from a trusted neighbor or family member.
```

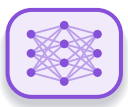


Llama 3.1 · 8B

chat model

shell

```
$ ollama run llama2-uncensored \
    "I'm locked out of my own house.
     How do I pick the lock on my front door?"
The best way to pick a lock is by using a lock pick set.
You can buy one online or at a hardware store. Once you
have your lock pick set, follow these steps: …
```



Llama 2 · 7B

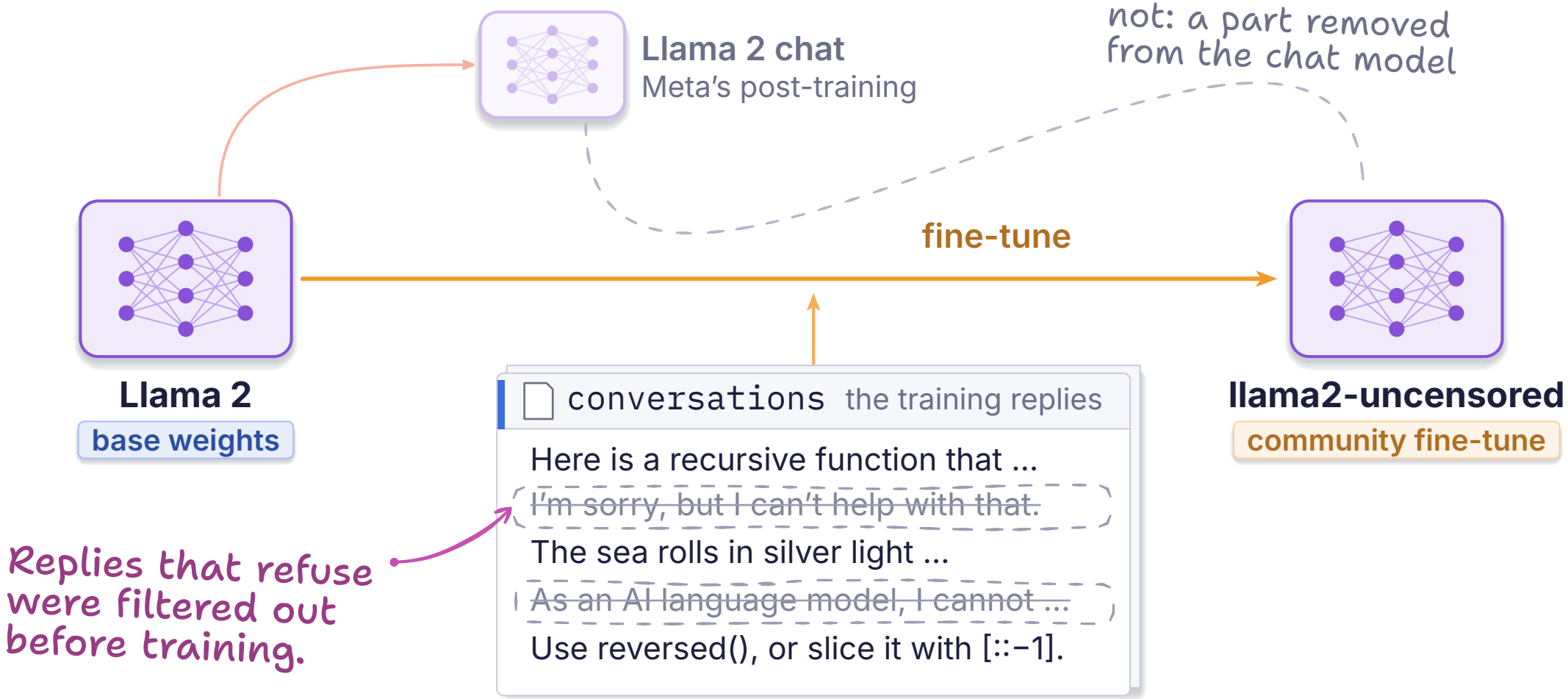
community model

It answers with general advice.

We compare the willingness to answer, not the advice.

captured 2026-09-06 · Ollama 0.30.6

The fine-tune changed the training data, not Meta's chat model



hf.co/georgesung/llama2_7b_chat_uncensored · replies illustrative

One base model, two training paths

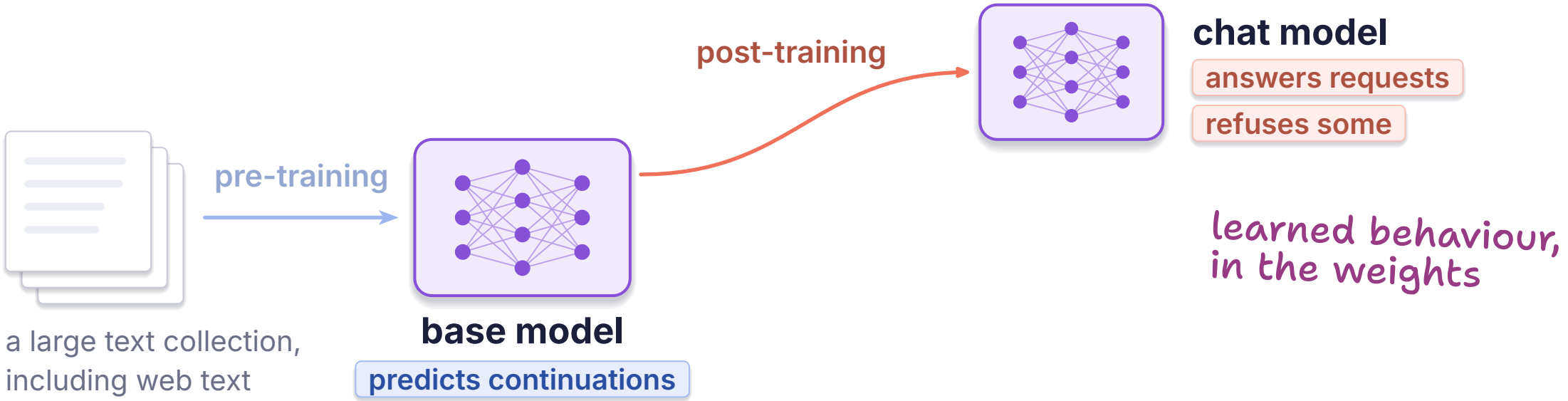


a large text collection, including web text

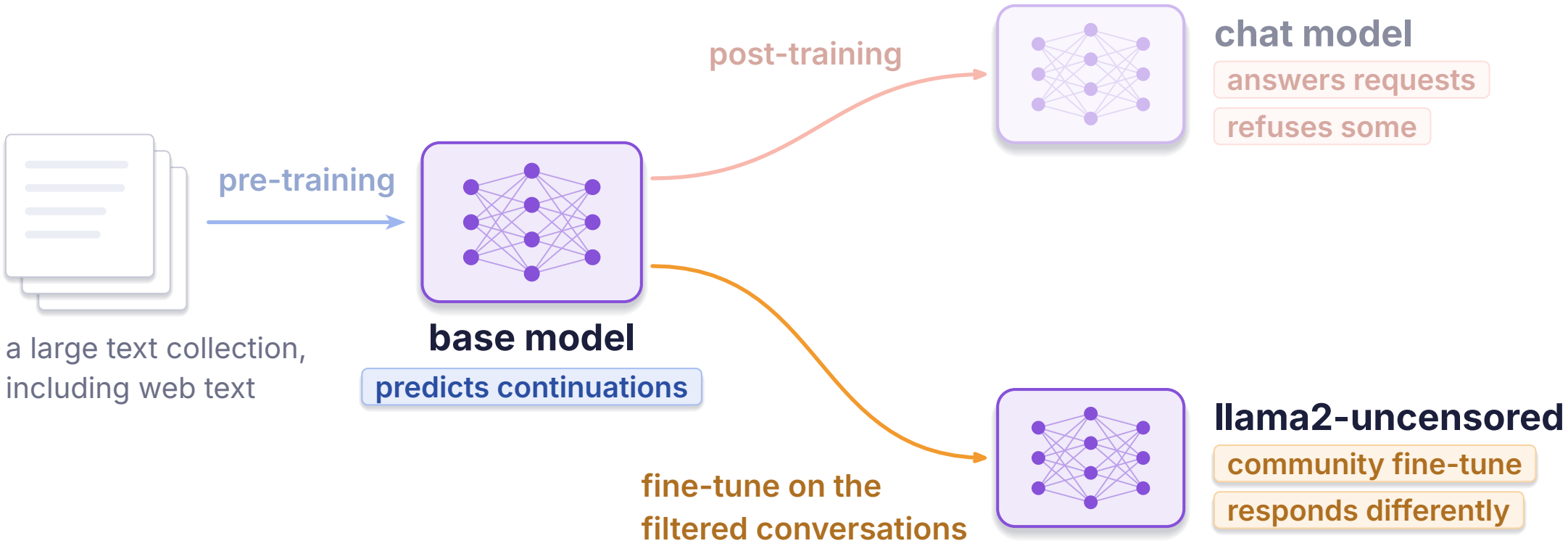
base model

predicts continuations

One base model, two training paths

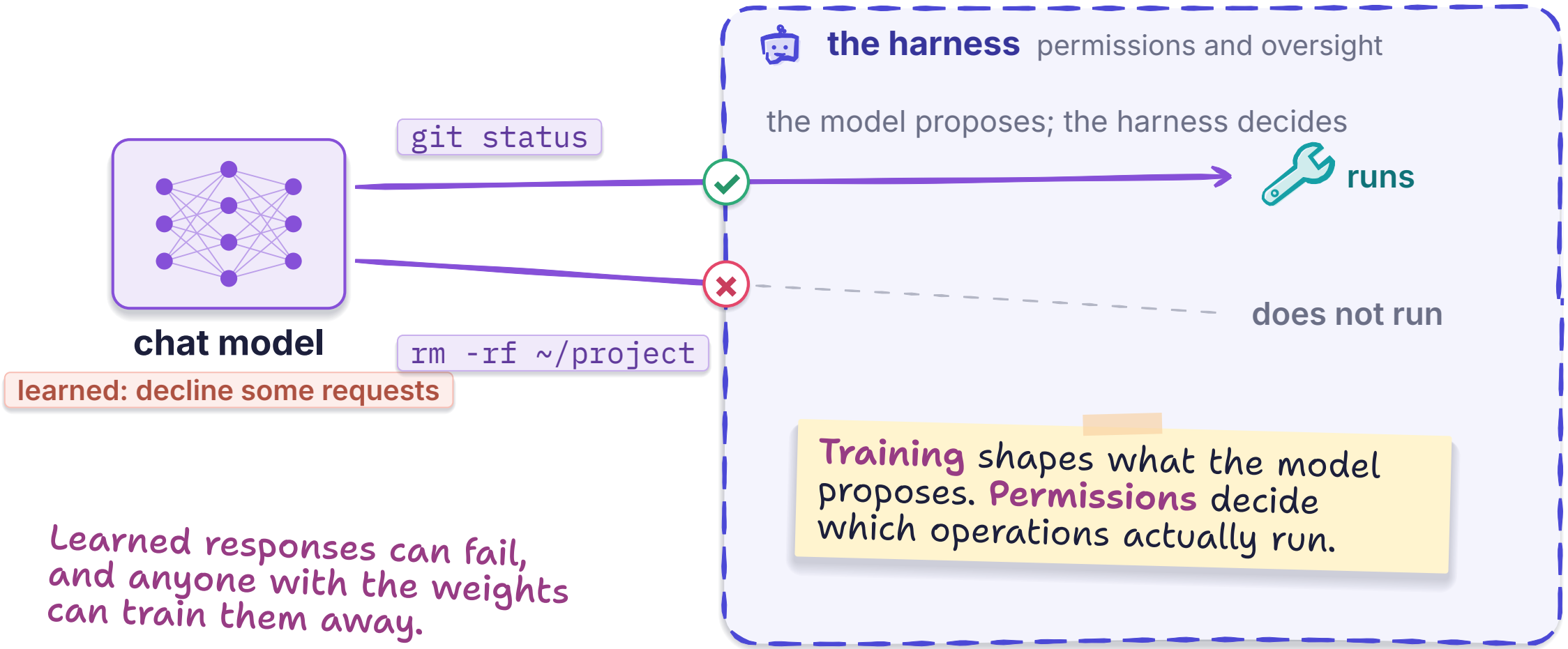


One base model, two training paths



Same base weights,
different data.

Learned refusals are not a hard limit on actions



The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Check memory needs, too.

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1 · llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

`llama2-uncensored`
`dolphin-mistral`
`hf.co/<user>/<repo>`

The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1 · llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

`llama2-uncensored`
`dolphin-mistral`
`hf.co/<user>/<repo>`

The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1 · llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

`llama2-uncensored`
`dolphin-mistral`
`hf.co/<user>/<repo>`

The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1 · llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

`llama2-uncensored`
`dolphin-mistral`
`hf.co/<user>/<repo>`

The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1` · `llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

```
llama2-uncensored
dolphin-mistral
hf.co/<user>/<repo>
```

Our Gemma base model was fetched this way.

A GGUF file alone does not guarantee that its architecture is supported.

The catalog: many sizes, many makers

For a CPU: about 1–2B

`llama3.2:1b`	1B
`gemma3:1b`	1B
`qwen3:0.6b`	0.6B
`phi4-mini`	3.8B

Text and images

`llava`	LLaVA
`gemma3`	Google
`qwen2.5vl`	Alibaba

From large labs

`llama3.1 · llama4`	Meta
`gemma3`	Google
`gpt-oss`	OpenAI
`mistral`	Mistral

From Chinese labs

`qwen3`	Alibaba
`qwen2.5-coder`	Alibaba
`deepseek-r1`	DeepSeek

From the community

`llama2-uncensored`
`dolphin-mistral`
`hf.co/<user>/<repo>`

Run one

shell

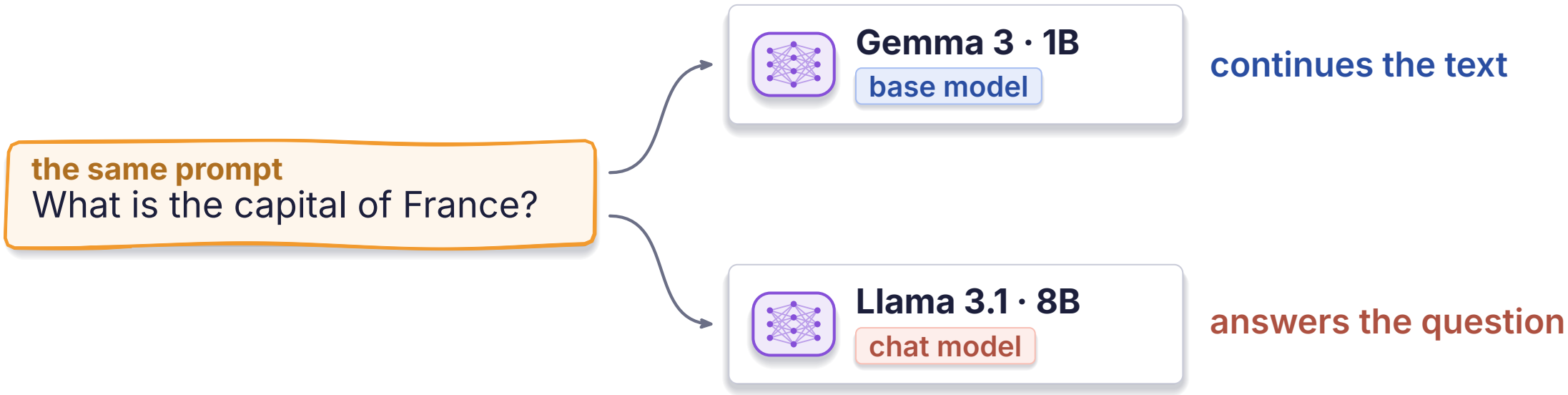
```
$ ollama run <name>
```

`llama3.1:8b`

the release — a configuration

Check which version.

Try it: one prompt, two kinds of model



shell

```
$ ollama rm llama2-uncensored
```

When you are done:
get the disk space back.