open models

artificial intelligence · prof. dr. nicolas meseth

1. what do you think?
2. offline
3. what a model is
4. open or closed

no internet, and it still answers.



image: ai-generated (gpt-image-2.5-flare)

part 1

what do you think?

when you chat with ChatGPT, where does the model run?

on my phone or laptop

in my browser

in a data centre

no idea

lifi.uber.space/q

when you chat with ChatGPT, where does the model run?

- on my phone or laptop
- in my browser
- in a data centre
- no idea

lifi.uber.space/q

when you chat with ChatGPT, where does the model run?

- on my phone or laptop
- in my browser
- in a data centre
- no idea



lifi.uber.space/q

how big is a language model that runs on a laptop?

a few megabytes

one to a few gigabytes

around 100 gigabytes

no idea



lifi.uber.space/q

how big is a language model that runs on a laptop?

- a few megabytes
- one to a few gigabytes
- around 100 gigabytes
- no idea

lifi.uber.space/q

how big is a language model that runs on a laptop?

- a few megabytes
- one to a few gigabytes
- around 100 gigabytes
- no idea



lifi.uber.space/q

part 2

offline

turn off wi-fi.

turn off wi-fi.

load gemma 3 1b in LM Studio.

turn off wi-fi.

load gemma 3 1b in LM Studio.

ask it: “what does nitrogen fertiliser do for a wheat plant? two sentences.”

LM Studio → my models → show in folder.

LM Studio → my models → show in folder.

how big is the file, and what is it called?

> a field of 4.5 ha needs 140 kg of nitrogen per ha. how much is that in total?

> a field of 4.5 ha needs 140 kg of nitrogen per ha. how much is that in total?

> what is the capital of lower saxony, and what is it known for?

- a field of 4.5 ha needs 140 kg of nitrogen per ha. how much is that in total?
- what is the capital of lower saxony, and what is it known for?
- translate into german: “the harvest was delayed by rain.”

> a field of 4.5 ha needs 140 kg of nitrogen per ha. how much is that in total?

> what is the capital of lower saxony, and what is it known for?

> translate into german: “the harvest was delayed by rain.”

ask gemma 3 1b first, then gemma 3 4b, and note the tokens per second below each answer.

how many tokens per second did gemma 3 4b reach on your laptop?

fewer than 5

5 to 20

20 to 50

more than 50

lifi.uber.space/q

how many tokens per second did gemma 3 4b reach on your laptop?

- fewer than 5
- 5 to 20
- 20 to 50
- more than 50

lifi.uber.space/q

1b or 4b?

	gemma 3 1b	gemma 3 4b
file		
speed		
quality		

1b or 4b?

	gemma 3 1b	gemma 3 4b
file	0.75 GB	2.4 GB
speed		
quality		

1b or 4b?

	gemma 3 1b	gemma 3 4b
file	0.75 GB	2.4 GB
speed	your tokens/s?	your tokens/s?
quality		

	gemma 3 1b	gemma 3 4b
file	0.75 GB	2.4 GB
speed	your tokens/s?	your tokens/s?
quality	which answers held?	which answers held?

if a model like GPT-6.1 Sol lay on your laptop, how big would the file be?

about 5 gigabytes

about 50 gigabytes

about 500 gigabytes

several terabytes

lifi.uber.space/q

if a model like GPT-6.1 Sol lay on your laptop, how big would the file be?

- about 5 gigabytes
- about 50 gigabytes
- about 500 gigabytes
- several terabytes

lifi.uber.space/q

if a model like GPT-6.1 Sol lay on your laptop, how big would the file be?

about 5 gigabytes

about 50 gigabytes

about 500 gigabytes

several terabytes

lifi.uber.space/q

part 3

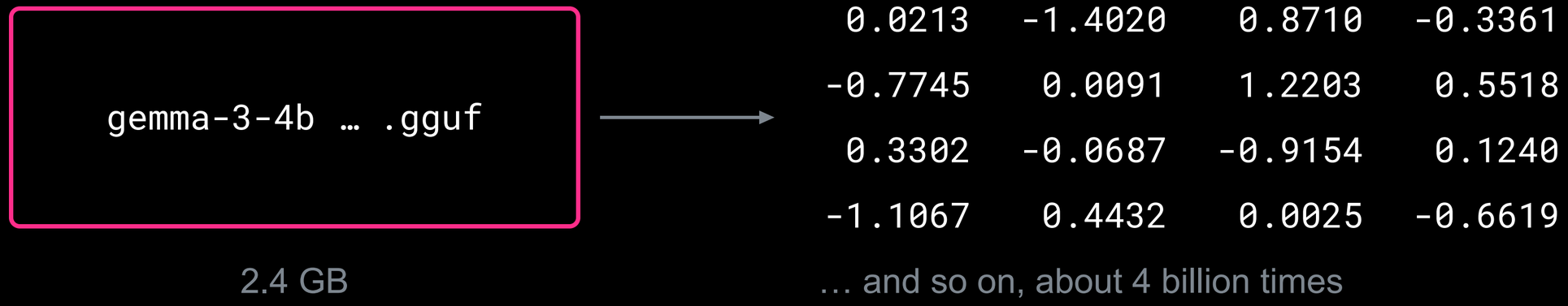
what a model is

a file full of numbers

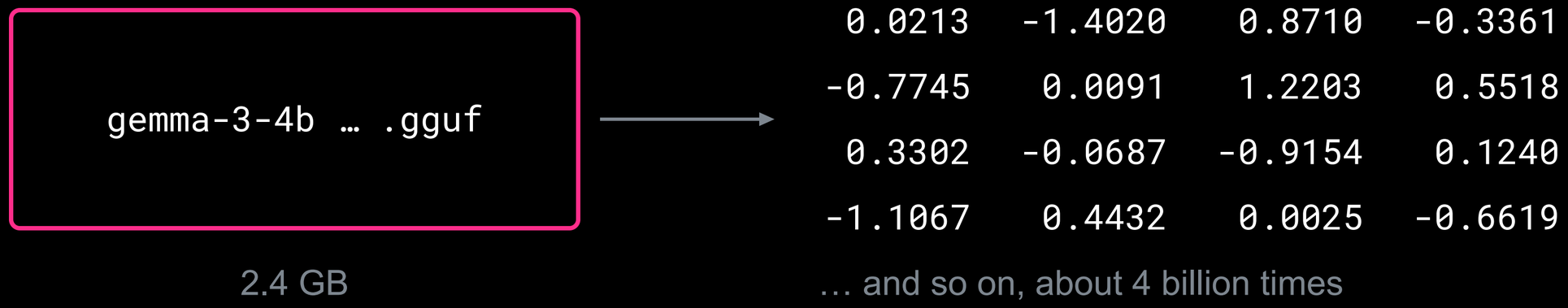
`gemma-3-4b … .gguf`

2.4 GB

a file full of numbers



a file full of numbers



4 billion parameters, also called weights

how many values fit into a few bits?

1 bit → 2 values

how many values fit into a few bits?

1 bit → 2 values

2 bits → 4 values

how many values fit into a few bits?

1 bit → 2 values

2 bits → 4 values

4 bits → 16 values

how many values fit into a few bits?

1 bit → 2 values

2 bits → 4 values

4 bits → 16 values

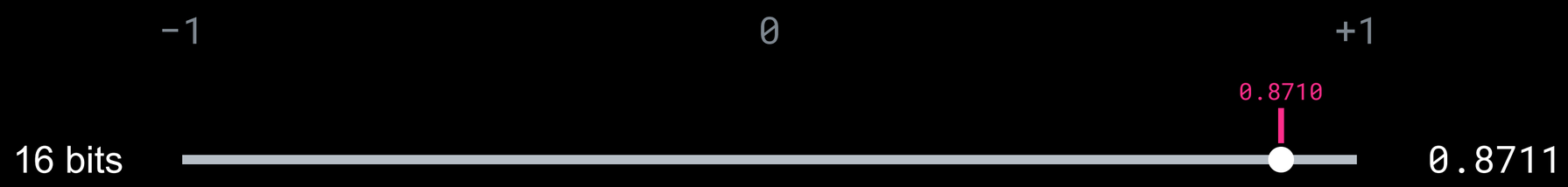
16 bits → 65,536 values

how many values fit into a few bits?

1 bit	→	2 values
2 bits	→	4 values
4 bits	→	16 values
16 bits	→	65,536 values

every extra bit doubles the number of values a weight can take.

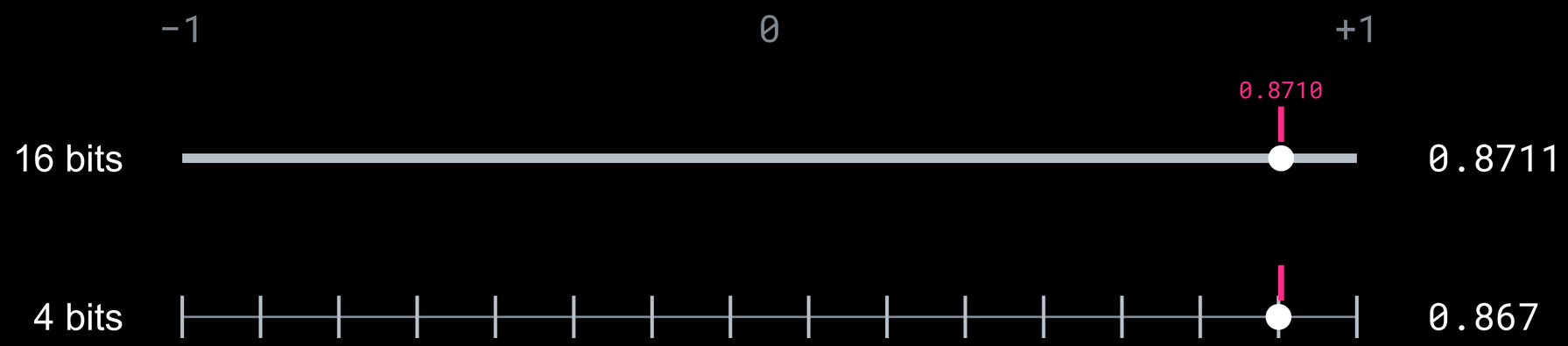
rounding one weight



simplified: 16 evenly spaced values between −1 and +1

what a model is

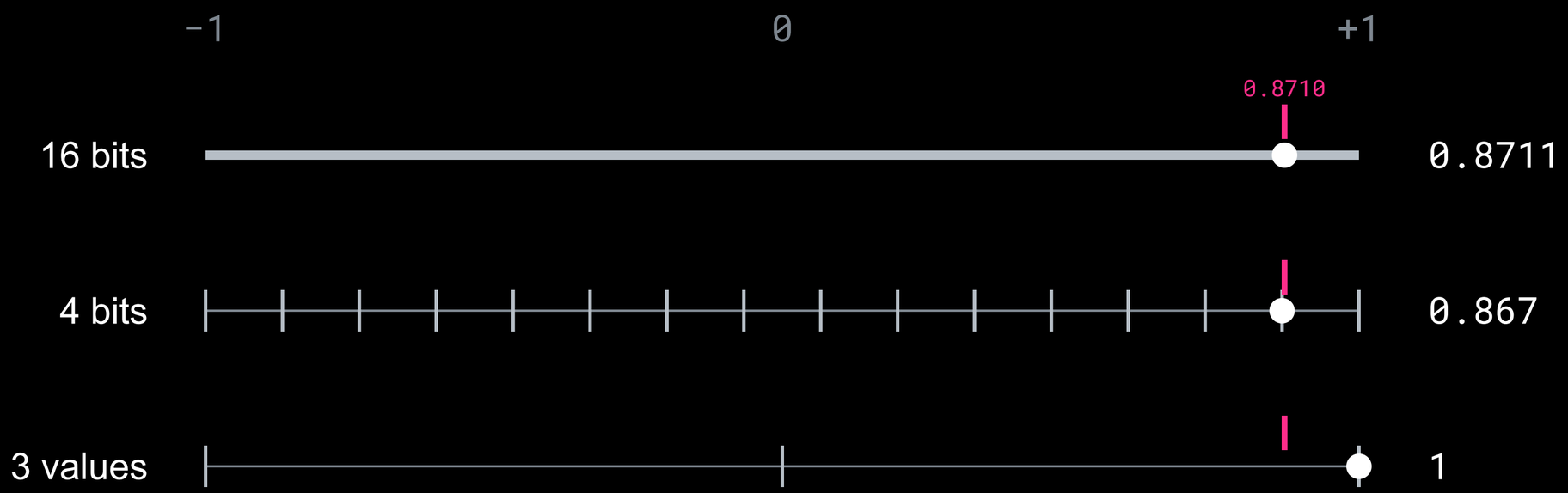
rounding one weight



simplified: 16 evenly spaced values between −1 and +1

what a model is

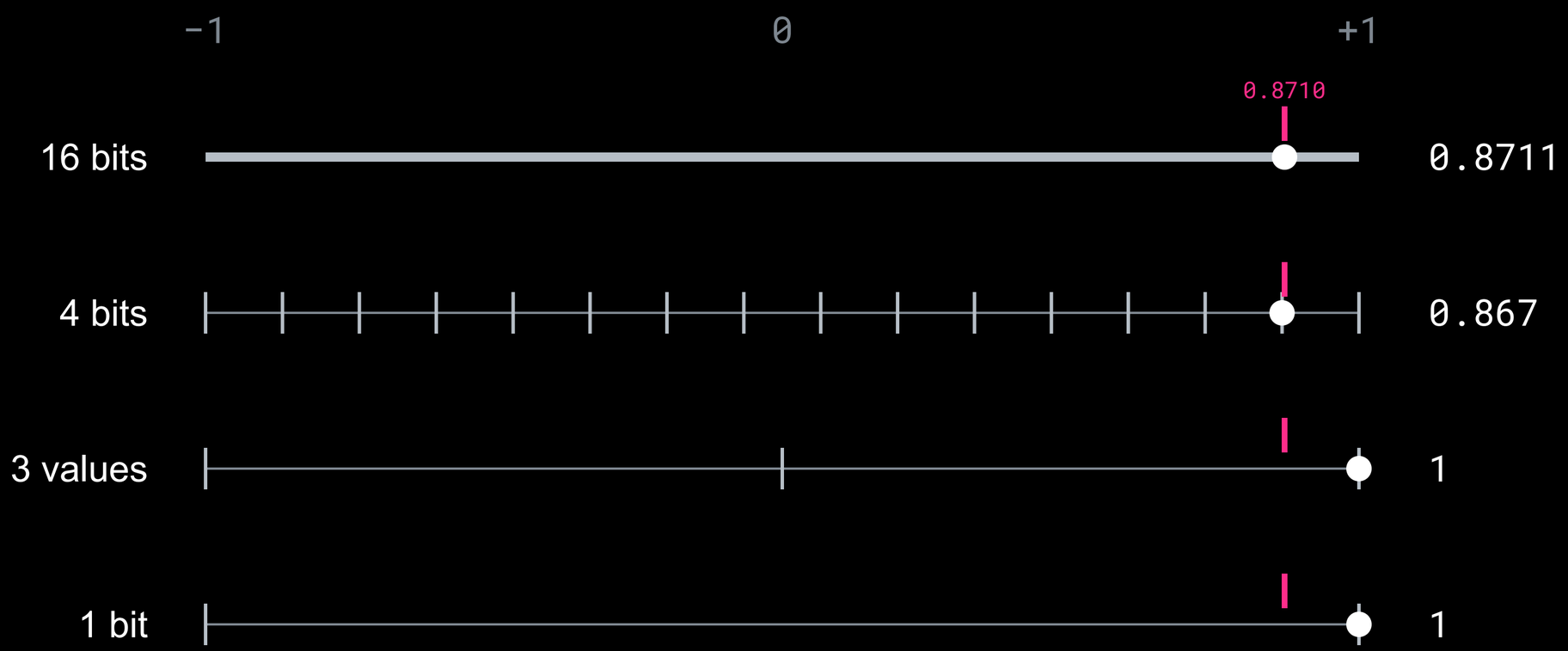
rounding one weight



simplified: 16 evenly spaced values between −1 and +1

what a model is

rounding one weight



simplified: 16 evenly spaced values between −1 and +1

what a model is

four weights, rounded

16 bits

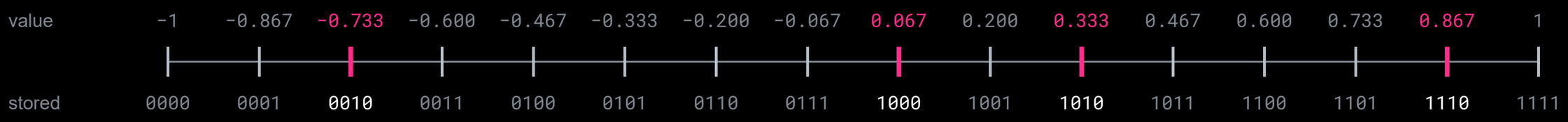
0.0213

-0.7744

0.3301

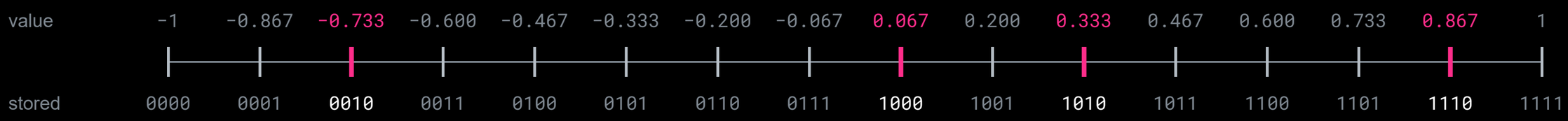
0.8711

four weights, rounded



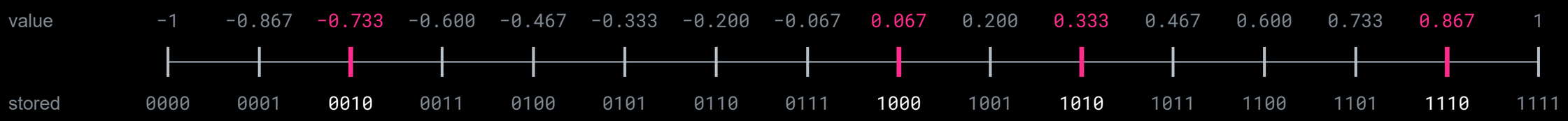
16 bits		4 bits
0.0213	0.067	1000
-0.7744	-0.733	0010
0.3301	0.333	1010
0.8711	0.867	1110

four weights, rounded



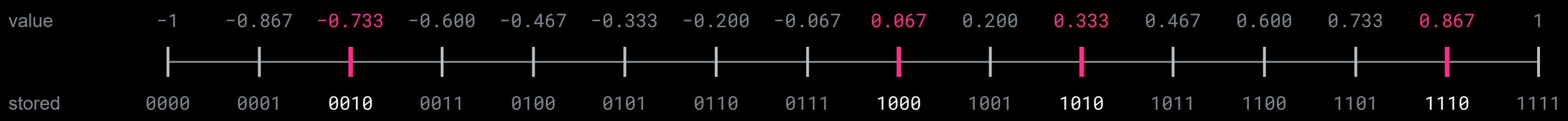
16 bits	4 bits		3 values
0.0213	0.067	1000	0
-0.7744	-0.733	0010	-1
0.3301	0.333	1010	0
0.8711	0.867	1110	1

four weights, rounded



16 bits	4 bits		3 values	binary
0.0213	0.067	1000	0	1
-0.7744	-0.733	0010	-1	-1
0.3301	0.333	1010	0	1
0.8711	0.867	1110	1	1

four weights, rounded



16 bits	4 bits		3 values	binary
0.0213	0.067	1000	0	1
-0.7744	-0.733	0010	-1	-1
0.3301	0.333	1010	0	1
0.8711	0.867	1110	1	1

the coarser the rounding, the smaller the file, and the further the model drifts from what it learned.

4 billion numbers × 16 bits = 8 GB

4 billion numbers × 16 bits = 8 GB

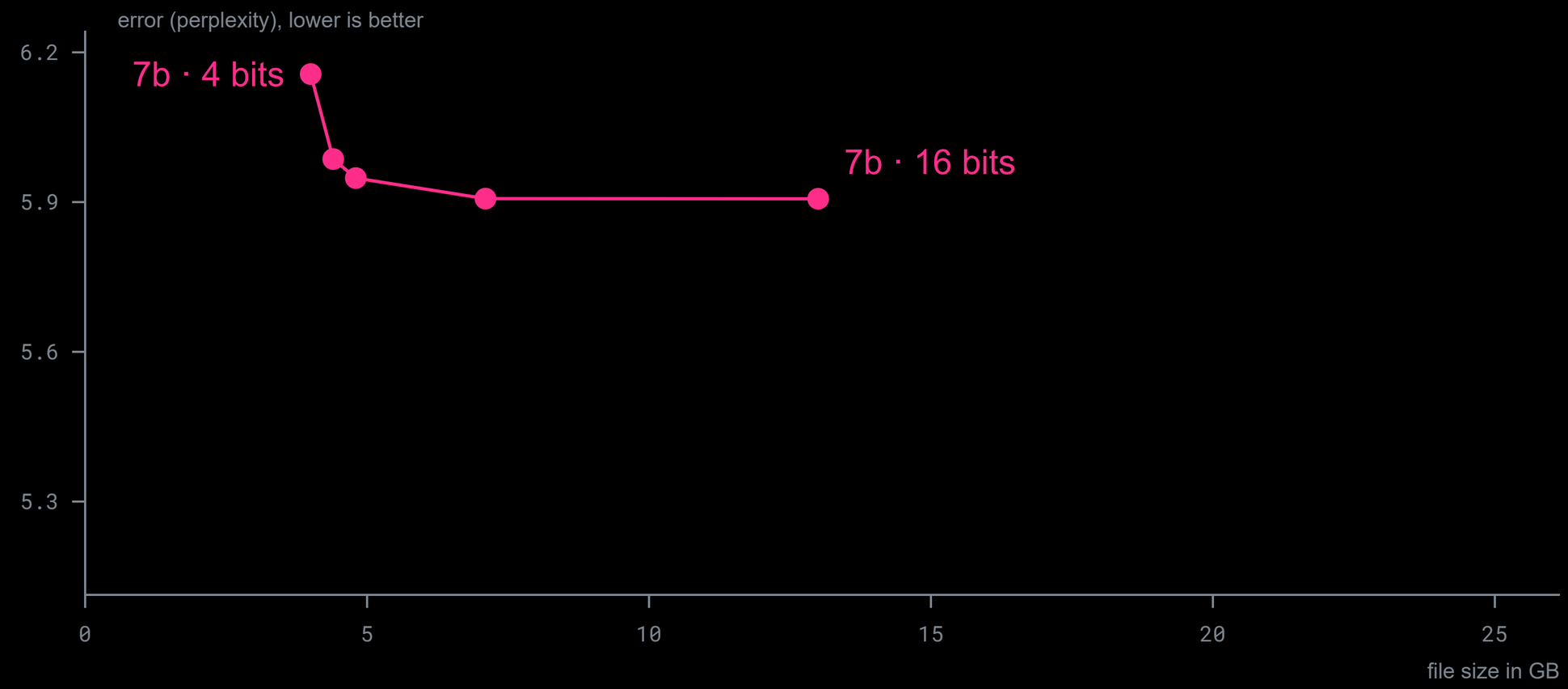
4 billion numbers × ~4.5 bits = 2.4 GB

4 billion numbers × 16 bits = 8 GB

4 billion numbers × ~4.5 bits = 2.4 GB

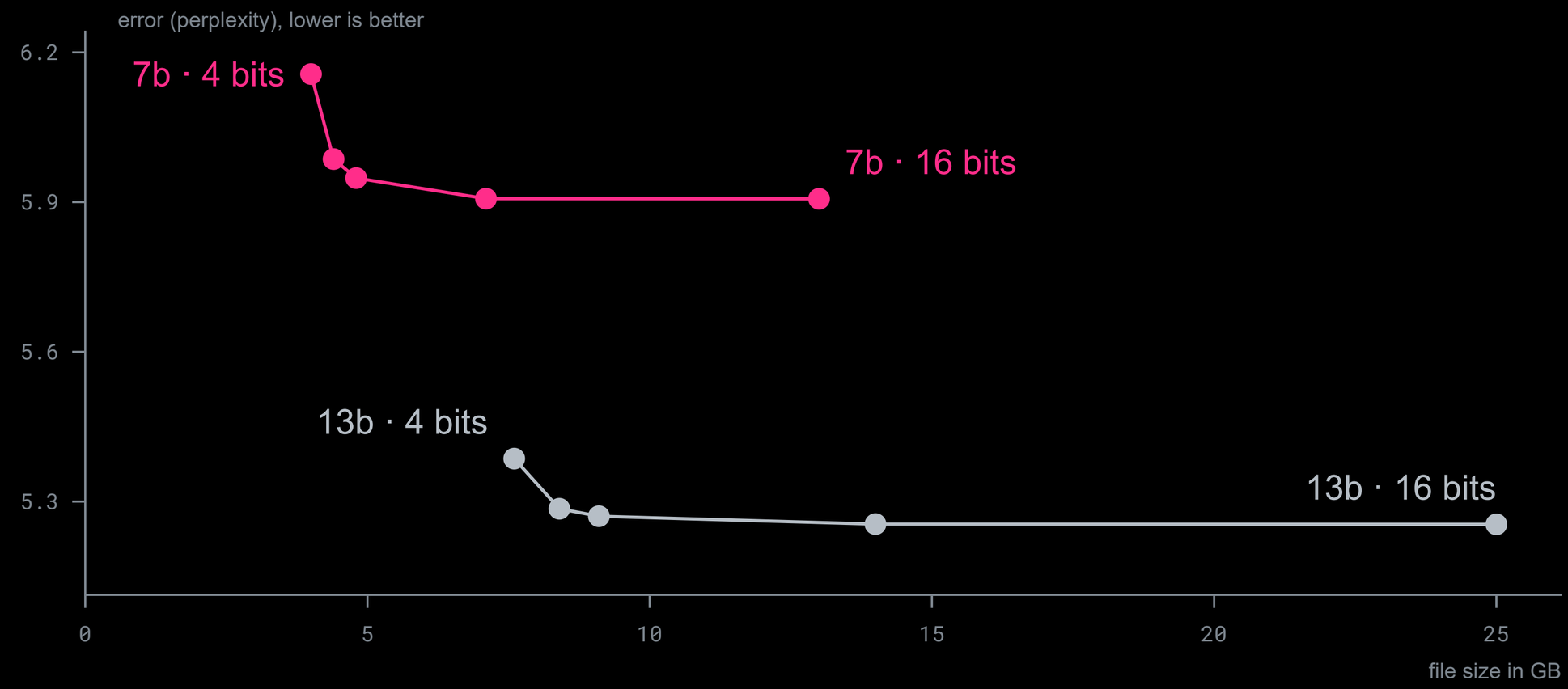
fewer bits per number: quantization.

rounding costs less than shrinking



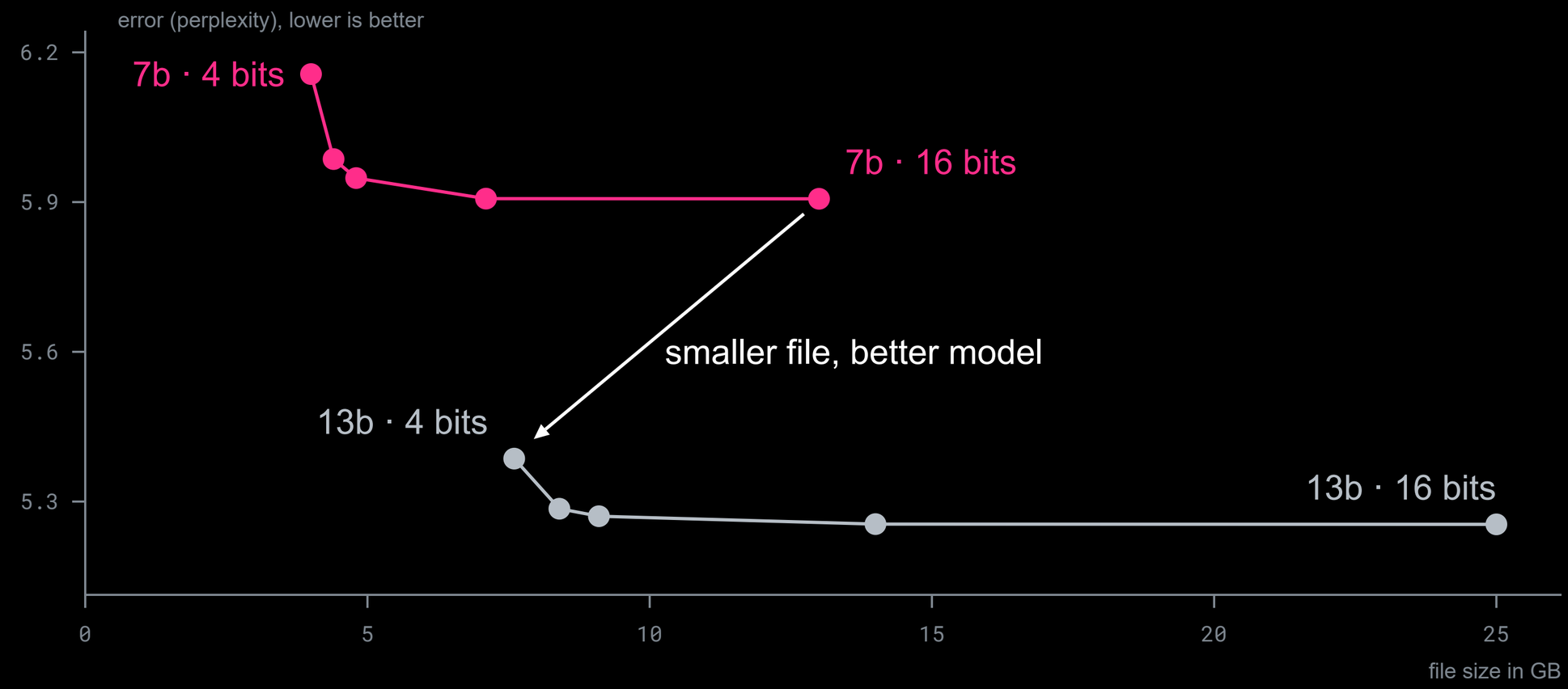
measured on LLaMA 7b and 13b: llama.cpp, 2023 · see also dettmers & zettlemoyer, the case for 4-bit precision, 2023

rounding costs less than shrinking



measured on LLaMA 7b and 13b: llama.cpp, 2023 · see also dettmers & zettlemoyer, the case for 4-bit precision, 2023

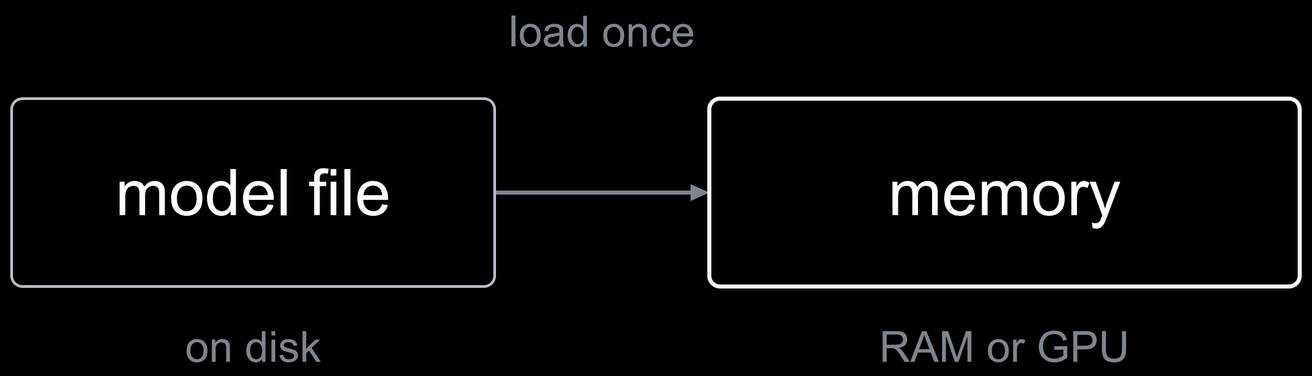
rounding costs less than shrinking

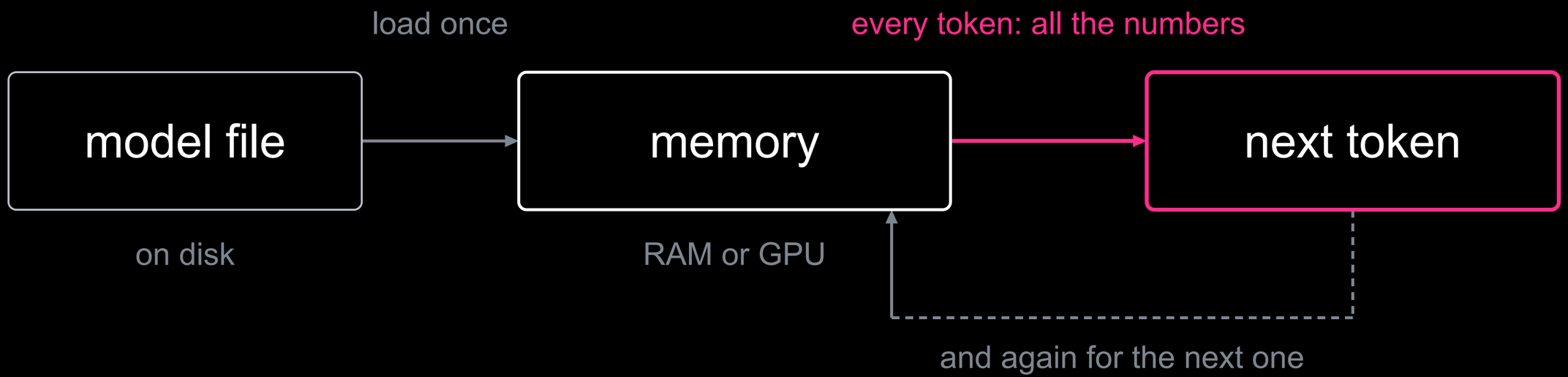


measured on LLaMA 7b and 13b: llama.cpp, 2023 · see also dettmers & zettlemoyer, the case for 4-bit precision, 2023

model file

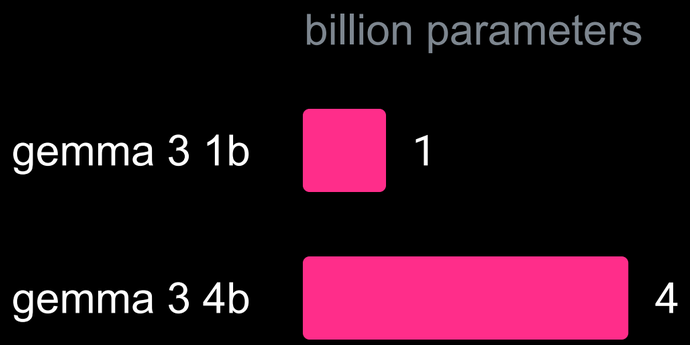
on disk


load once
model file
memory
on disk
RAM or GPU



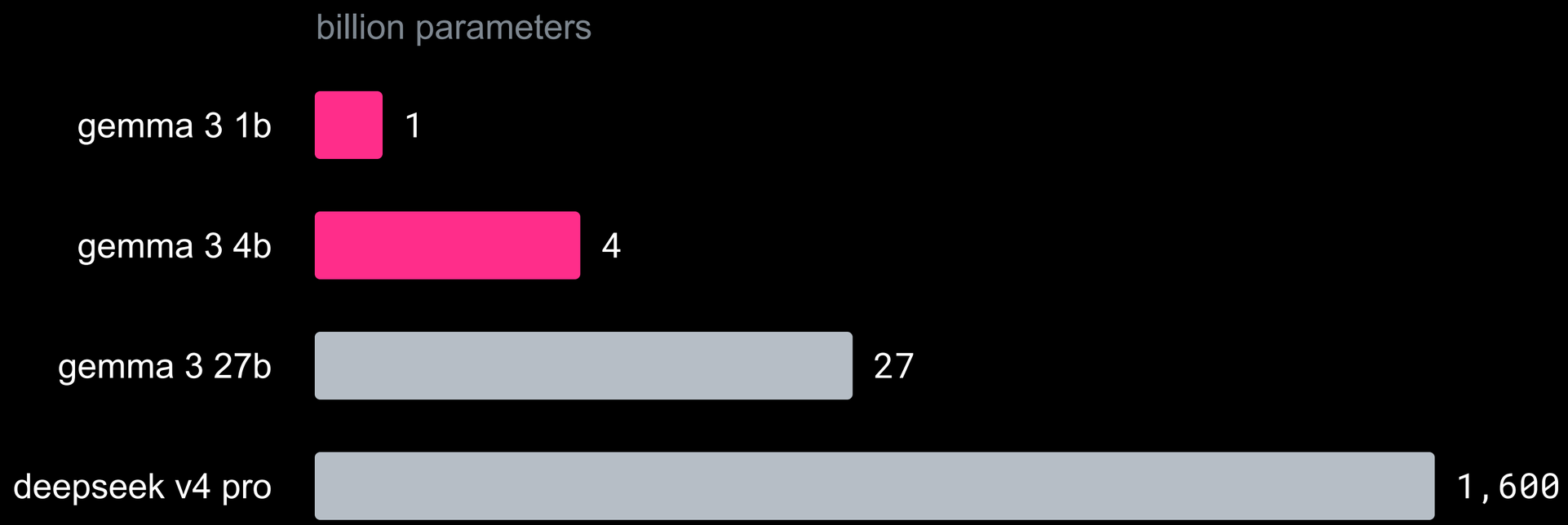
bigger models are slower, because every token passes through more numbers.

how big is big?



parameters as published by the providers; log scale

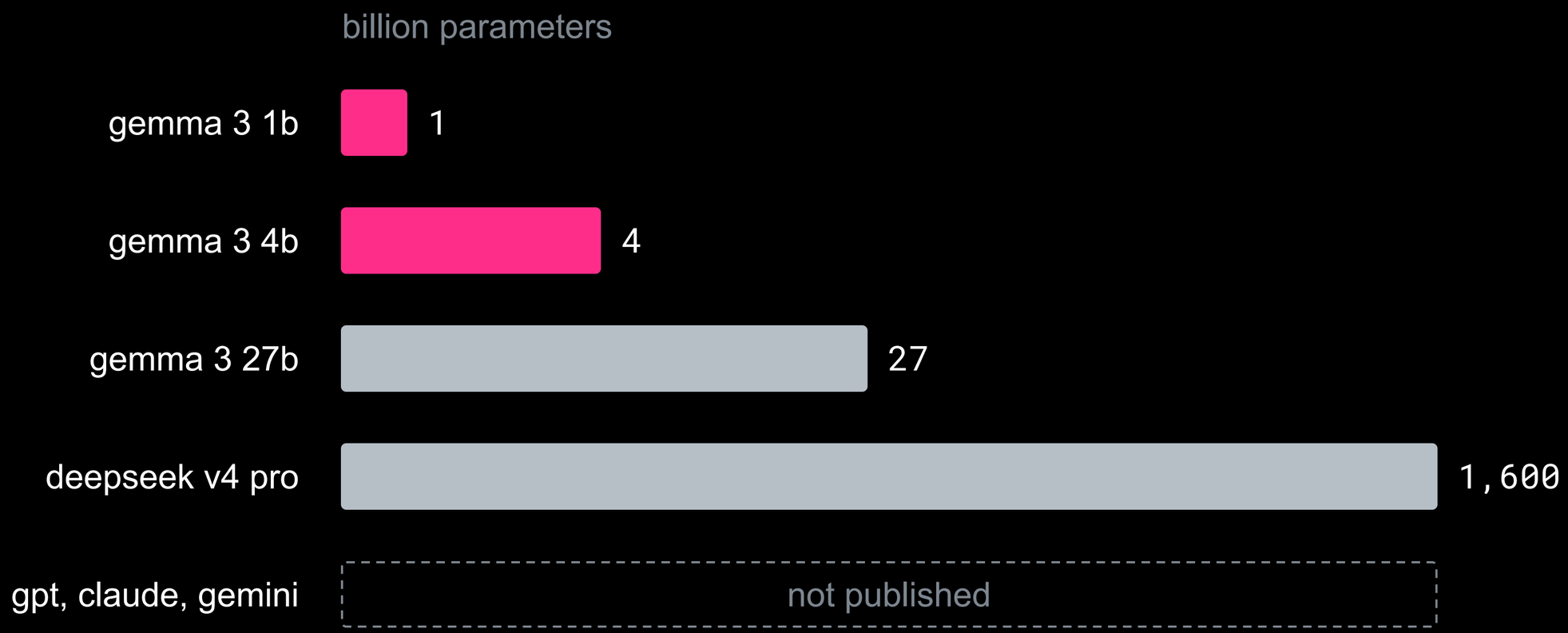
how big is big?



parameters as published by the providers; log scale

what a model is

how big is big?



parameters as published by the providers; log scale

the same harness can talk to a model on your laptop.

see also: the harness

part 4

open or closed

open

download it, run it, change it.

gemma, qwen, llama, mistral, deepseek, gpt-oss

open

download it, run it, change it.

gemma, qwen, llama, mistral, deepseek, gpt-oss

closed

use it as a service.

gpt, claude, gemini

open

download it, run it, change it.

gemma, qwen, llama, mistral, deepseek, gpt-oss

closed

use it as a service.

gpt, claude, gemini

open weights is not the same as open source: the training data usually stays secret.

would you paste a farm's customer list into a cloud chatbot?

yes

only with a contract

only into a local model

never



lifi.uber.space/q

would you paste a farm's customer list into a cloud chatbot?

- yes
- only with a contract
- only into a local model
- never

lifi.uber.space/q

your data must not leave the building.

your data must not leave the building.

you need it offline, in a barn, on a tractor or on your phone.

on a phone: google ai edge gallery, an app that runs small gemma models offline on android and ios

your data must not leave the building.

you need it offline, in a barn, on a tractor or on your phone.

you use it so much that tokens get expensive.

on a phone: google ai edge gallery, an app that runs small gemma models offline on android and ios

your data must not leave the building.

you need it offline, in a barn, on a tractor or on your phone.

you use it so much that tokens get expensive.

and you accept a smaller, slower model in return.

on a phone: google ai edge gallery, an app that runs small gemma models offline on android and ios

a language model is a file full of numbers that runs on your own laptop without the internet; its size decides how good, how fast and how hungry it is.

what surprised you about the size, speed or quality of the local model?

what surprised you about the size, speed or quality of the local model?

which data would you trust a local model with, but not a cloud service?

what surprised you about the size, speed or quality of the local model?

which data would you trust a local model with, but not a cloud service?

when should a farm or a food company run its own model instead of renting one?