

Run AI on **your own** computer

Local language models with Ollama and Olmo 3

THE QUESTION

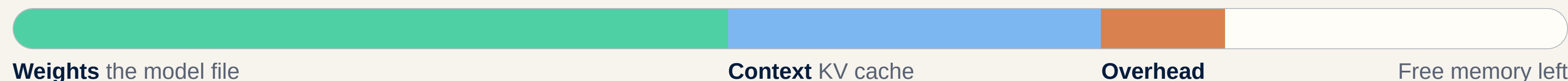
Can a capable language model run on hardware you already own?

THE PROMISE

Your prompts stay on your machine, and you choose the model.

OUR JOB

Pick a model that fits, install Ollama, and run Olmo 3.



Why run a model **locally**?

	Hosted model	Local model
Where your prompt goes	Sent to a provider’s servers	Stays on your machine
What it costs	Per token or by subscription	Hardware and power you already have
Works offline	Needs a connection	Yes, after the download
Who is in control	The provider can change or retire the model	You choose the version and keep the files
Capability and speed	The largest models, fast	Smaller models, limited by your memory

Local is not automatically private. Check for cloud features, network settings, and where the model came from.

What is inside a local model?

LAYER 1

Weights

Billions of numbers saved in one file. Olmo 3 7B is about 4.5 GB, and 32B is about 19 GB.

+

LAYER 2

Runtime

The program that loads the weights and does the math. Ollama builds on engines such as llama.cpp.

+

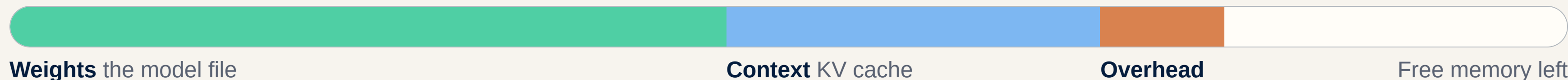
LAYER 3

Hardware

Memory holds the model, and memory bandwidth sets how fast words appear.

Memory $\approx$ parameters $\times$ bits per weight $\div$ 8 + context + overhead

Quantization stores each weight in fewer bits. Going from 16-bit to 4-bit makes the file about four times smaller, with a small quality cost.



What to watch for in the demo

How local models work

- Memory is the limit, not just compute
- Memory bandwidth sets tokens per second
- Context length costs memory (the KV cache)
- Dense or mixture-of-experts models
- Quantization: 4-bit is the common default

Ollama and Olmo 3

- run, pull, ps and stop
- Olmo 3: 7B is about 4.5 GB, 32B about 19 GB
- Instruct answers directly, Think shows its reasoning
- Cloud features can be switched off

Predict: can a laptop with 16 GB of memory run Olmo 3 32B Think? What about 7B Instruct?

Write down your answer. We'll check it in the demo.