

## FLASH CARDS

# Run AI on **your own computer**

Terminology and quiz cards for the lesson on local language models with Ollama and Olmo 3.

## What is in this set

---

40 cards: **24 term cards** (term on the front, definition on the back) and **16 quiz cards** (question on the front, answer on the back). Bracketed numbers on the backs, like [1], match the reference list on the last page and in the instructor guide.

## How to print

---

1. Print double-sided on US Letter at 100% scale (no "fit to page").
2. Choose **flip on the long edge**. The backs are arranged so each answer lands behind its card.
3. Cut along the dashed lines. Each card is 3.75 by 2.5 inches.

## How to use them

---

- **Self-test:** read the front, answer aloud, flip to check.
- **Pairs:** one reads the front, the other answers. Swap every five cards.
- **Boxes:** keep a "missed" pile and a "got it" pile. Re-test the missed pile after a break, then again the next day.
- **Warm-up:** deal a few quiz cards at the start of the next session.

Facts checked October 27, 2026. Tools and models change quickly, so confirm details against the current documentation. These materials are dedicated to the public domain under CC0 1.0 and shared for the public good. The PrivacySafe and PrivacySafe Foundation names and logos are trademarks and are not included in that dedication.

TERM

1 / 40

# Local model

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

2 / 40

# Weights (parameters)

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

3 / 40

# Inference

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

4 / 40

# Ollama

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

5 / 40

# Runtime (engine)

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

6 / 40

# GGUF

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

7 / 40

# Quantization

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

8 / 40

# Token

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

2 / 40

The billions of learned numbers that make up a model. Olmo 3 comes in 7B and 32B sizes.

[\[7\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

1 / 40

A model whose weights you download and run on your own hardware, so your prompts need not leave your machine.

[\[1\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

4 / 40

An open-source tool (MIT license) for downloading and running models locally, with a command line and a local API.

[\[3\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

3 / 40

Running a trained model to produce output. Training is how it learned; inference is using it.

**General**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

6 / 40

A file format for model weights used by llama.cpp and Ollama. Ollama can also import Safetensors.

[\[8\]](#) [\[10\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

5 / 40

The program that loads weights and does the math, such as llama.cpp. Tools like Ollama and LM Studio build on engines like it.

[\[8\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

8 / 40

A chunk of text, often part of a word, that the model reads and writes. English averages roughly three quarters of a word per token.

**General**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

7 / 40

Storing weights in fewer bits, such as 4-bit instead of 16-bit. It makes the file about four times smaller with a small quality cost.

[\[4\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

9 / 40

## Context window

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

10 / 40

## KV cache

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

11 / 40

## VRAM

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

12 / 40

## Unified memory

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

13 / 40

## Memory bandwidth

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

14 / 40

## Tokens per second

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

15 / 40

## Dense model

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

16 / 40

## Mixture of experts

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

Memory that stores attention keys and values for every token so far. It grows with context length and can be quantized.

[1]

How many tokens the model can consider at once. Ollama's default is 4096. Olmo 3 supports up to 65,536.

[1] [7]

One memory pool shared by the CPU and GPU, as on Apple Silicon. It lets larger models fit.

[2]

Video memory on a GPU. If the model fits there, it runs fastest.

[2]

Generation speed. About 5 is near reading speed. 20 or more feels interactive.

Planner

How fast memory can be read, in GB per second. Each token reads the weights, so bandwidth largely sets tokens per second.

General

A model that activates only some experts per token. Memory holds all parameters; speed follows the active ones.

General

A model that uses all its parameters for every token. Memory and speed both track its total size.

General

TERM

17 / 40

## Instruct model

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

18 / 40

## Think model

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

19 / 40

## Open weights

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

20 / 40

## Fully open model

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

21 / 40

## Model card

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

22 / 40

## ollama ps

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

23 / 40

## OLLAMA\_NO\_CLOUD

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

TERM

24 / 40

## localhost (127.0.0.1)

Say it in your own words, then flip.

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

18 / 40

Trained to reason step by step before answering, like Olmo 3 Think. Stronger on hard problems, slower to reply.

[\[7\]](#) [\[4\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

17 / 40

Tuned to follow instructions and answer directly, like Olmo 3 7B Instruct.

[\[4\]](#) [\[7\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

20 / 40

Weights, training code and detailed data information are all released. Ai2 releases these for Olmo 3.

[\[13\]](#) [\[5\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

19 / 40

The parameters can be downloaded, but code or data may be withheld and licenses can add conditions.

[\[13\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

22 / 40

Lists loaded models with their size and PROCESSOR: 100% GPU, 100% CPU, or a split.

[\[1\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

21 / 40

The page describing a model: size, context, license, training data and limits. Read it before you rely on a model.

[\[7\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

24 / 40

Your own machine. Ollama listens on 127.0.0.1 port 11434 by default, so other devices cannot reach it unless you change OLLAMA\_HOST.

[\[1\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**TERM**

23 / 40

Environment variable set to 1 that turns off Ollama's cloud models and web search for local-only use.

[\[1\]](#)

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

25 / 40

**Roughly how much memory do the weights of a 7B model need at 4-bit?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

26 / 40

**Which three things add up to the memory a model needs?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

27 / 40

**What is Ollama's default context length?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

28 / 40

**Which command shows where a model loaded?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

29 / 40

**How do you free a model's memory right now?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

30 / 40

**What does 100% CPU in ollama ps tell you?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

31 / 40

**Why can a 16 GB laptop not run Olmo 3 32B Think?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

32 / 40

**Instruct versus Think: what is the difference?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

26 / 40

Weights, the context cache, and overhead.

General

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

25 / 40

About  $7.3 \times 4.85 / 8$ , or 4.4 GB. The Olmo 3 7B download is about 4.5 GB.

[4]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

28 / 40

**ollama ps.** The PROCESSOR column shows GPU, CPU or a split.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

27 / 40

4096 tokens. Change it with OLLAMA\_CONTEXT\_LENGTH.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

30 / 40

The model is running in system memory, not on a GPU. Expect it to be slower.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

29 / 40

**ollama stop** followed by the model name. Otherwise models stay loaded for 5 minutes.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

32 / 40

Instruct answers directly. Think reasons step by step first, so replies are longer and slower.

[7]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

31 / 40

It needs about 21.5 GB at 4096 tokens of context: about 19 GB of weights plus cache and overhead.

[4]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

33 / 40

**Why does memory bandwidth matter for speed?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

34 / 40

**For a mixture-of-experts model, which number sets memory and which sets speed?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

35 / 40

**What happens if a model does not fit in video memory?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

36 / 40

**Name two reasons local AI may be less private than it looks.**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

37 / 40

**Open weights versus fully open: what is the difference?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

38 / 40

**Which environment variable changes the default context length?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

39 / 40

**What does quantization trade away?**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

QUIZ

40 / 40

**Name three of the eight questions for choosing a local model.**

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

34 / 40

Total parameters set memory. Active parameters set speed.

General

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

33 / 40

Every token reads the weights, so tokens per second is roughly bandwidth divided by the bytes read per token.

General

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

36 / 40

Cloud models or web search are enabled; the server is exposed to the network; the model came from an untrusted source.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

35 / 40

Ollama splits it between the GPU and system RAM. It works but is slower, and ollama ps shows the split.

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

38 / 40

**OLLAMA\_CONTEXT\_LENGTH.**

[1]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

37 / 40

Open weights share the parameters. Fully open also shares the code and detailed data information.

[13]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

40 / 40

For example: how much memory can it use; do the weights fit; how much context do I need; what is the license; is anything leaving the machine.

[1] [13]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

**QUIZ**

39 / 40

A little quality, in exchange for a much smaller file and less memory.

[4]

Local AI with Ollama

PrivacySafe Foundation | CC0 1.0

## Reference list

Links were checked on October 27, 2026.

| #    | SOURCE                                                                                                                                                                         | WHAT IT SUPPORTS                                                                                                                                                                           |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [1]  | Ollama documentation, FAQ<br><a href="https://docs.ollama.com/faq">https://docs.ollama.com/faq</a>                                                                             | Default 4096-token context; ollama ps and the PROCESSOR column; keep-alive of 5 minutes; OLLAMA_HOST binding to 127.0.0.1:11434; OLLAMA_NO_CLOUD; KV cache types; model storage; updating. |
| [2]  | Ollama documentation, Hardware support<br><a href="https://docs.ollama.com/gpu">https://docs.ollama.com/gpu</a>                                                                | NVIDIA, AMD (ROCm), Apple (Metal) and experimental Vulkan support.                                                                                                                         |
| [3]  | Ollama on GitHub<br><a href="https://github.com/ollama/ollama">https://github.com/ollama/ollama</a>                                                                            | The project, the MIT license, and the Linux install script.                                                                                                                                |
| [4]  | Ollama library, olmo-3<br><a href="https://ollama.com/library/olmo-3">https://ollama.com/library/olmo-3</a>                                                                    | Olmo 3 tags and commands, download sizes (about 4.5 GB for 7B and 19 GB for 32B), 64K context, Instruct and Think variants.                                                                |
| [5]  | Ai2 documentation, Introduction<br><a href="https://docs.allenai.org/">https://docs.allenai.org/</a>                                                                           | Ai2's description of Olmo 3 and of releasing training code, data, weights and recipes.                                                                                                     |
| [6]  | Ai2, Olmo 3 technical report page<br><a href="https://allenai.org/papers/olmo3">https://allenai.org/papers/olmo3</a>                                                           | The Olmo 3 paper, referenced from the model cards.                                                                                                                                         |
| [7]  | Ai2 on Hugging Face (Olmo 3 model cards)<br><a href="https://huggingface.co/allenai">https://huggingface.co/allenai</a>                                                        | Model sizes, architecture table, 65,536-token context, Dolma 3 and Dolci data, Apache 2.0 license.                                                                                         |
| [8]  | llama.cpp on GitHub<br><a href="https://github.com/ggml-org/llama.cpp">https://github.com/ggml-org/llama.cpp</a>                                                               | The C++ inference engine and the GGUF format that many local tools build on.                                                                                                               |
| [9]  | Ollama documentation, OpenAI compatibility<br><a href="https://docs.ollama.com/api/openai-compatibility">https://docs.ollama.com/api/openai-compatibility</a>                  | The /v1 endpoint that lets OpenAI-style clients talk to a local Ollama server.                                                                                                             |
| [10] | Ollama documentation, Importing models<br><a href="https://docs.ollama.com/import">https://docs.ollama.com/import</a>                                                          | Model formats Ollama can import (GGUF and Safetensors).                                                                                                                                    |
| [11] | vLLM documentation<br><a href="https://docs.vllm.ai/">https://docs.vllm.ai/</a>                                                                                                | A serving system for running models for many users.                                                                                                                                        |
| [12] | LM Studio<br><a href="https://lmstudio.ai/">https://lmstudio.ai/</a>                                                                                                           | A desktop application with a model browser. Free to use; not open source.                                                                                                                  |
| [13] | Open Source Initiative, The Open Source AI Definition<br><a href="https://opensource.org/ai/open-source-ai-definition">https://opensource.org/ai/open-source-ai-definition</a> | What open source means for AI: data information, code and parameters, plus the four freedoms.                                                                                              |
| [14] | Meta, Llama license<br><a href="https://www.llama.com/llama3/license/">https://www.llama.com/llama3/license/</a>                                                               | An example of open weights under a custom license.                                                                                                                                         |
| [15] | Google, Gemma terms of use<br><a href="https://ai.google.dev/gemma/terms">https://ai.google.dev/gemma/terms</a>                                                                | An example of open weights under custom terms.                                                                                                                                             |
| [16] | Creative Commons, CC0 1.0 Universal<br><a href="https://creativecommons.org/publicdomain/zero/1.0/">https://creativecommons.org/publicdomain/zero/1.0/</a>                     | The public-domain dedication that covers these lesson materials.                                                                                                                           |