

Gemma: Google's artificial intelligence on your computer

A taste of open models, which run locally, with no subscription and no internet

Gianluca Bonomo · www.gianlu.it · info@gianlu.it



Contents

How to use this handout	3
1. Gemini or Gemma?	3
2. Why run it locally	4
3. The features of the Gemma families	5
3.1 Gemma 4 (April 2026)	5
3.2 Gemma 3 (the previous generation)	5
4. How it is launched, in broad terms	6
5. When it makes sense (and when not)	6
Hands-on exercise	8
Questions that come up in class	9
Glossary	9
Further reading	9

How to use this handout

This handout gathers, in written and ordered form, everything covered in the lesson. You do not need to memorise it: it is there so you can find things again calmly, later, when the lesson is over and a question is left hanging.

GOAL

Understand the difference between Gemini and Gemma, what it means to run a model locally and when it makes sense, without going into the details of installation.

What you will be able to do

- Explain the difference between Gemini (in the cloud) and Gemma (open models).
- List the advantages and limits of a model that runs on your own computer.
- Recognise the main features of the Gemma 3 and Gemma 4 families.
- Know what Ollama is and, in broad terms, how a model is launched.

1. Gemini or Gemma?

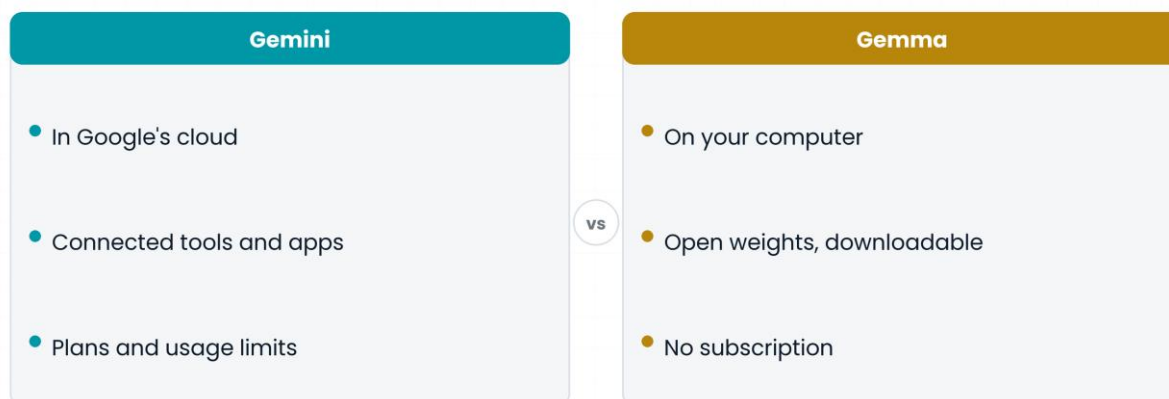
Gemini is a restaurant: you order, they serve you, you pay (or not). Gemma is the recipe: you take it home and cook yourself.

So far we have talked about **Gemini**, which lives in Google's computers: you write, the request travels over the internet, the answer comes back. **Gemma** is another family of Google models, with a fundamental difference: it is **open-weight**. The «weights» are the numerical heart of the model, and when they are open you can **download** them and run them on your own computer.

IN PLAIN WORDS

Gemini you use from the cloud. Gemma you bring home. They are relatives, not the same thing: Gemma is not «Gemini in a local version».

The restaurant and the recipe



Gemini and Gemma compared

2. Why run it locally

Running a model on your own computer is called **locally**. It has real advantages and equally real limits.

Advantages	Limits
Privacy: texts do not leave the computer.	Hardware: you need memory and, better, a good graphics card.
Works offline, after the download.	Quality: a small model is less capable than a big one in the cloud.
No subscription and no plan usage limits.	No connected tools: no Gmail, Drive, web search as in Gemini.
Freedom: you can try it, adapt it, use it in your own programs.	Maintenance: updating, downloading and managing files is up to you.

Locally: what you gain and what you lose.

THE WEIGHT OF THE FILES

Models are big files: from a few hundred megabytes to tens of gigabytes, depending on size. Before downloading, check the free space and the size shown on the model's official page.

3. The features of the Gemma families

Here is a taste of what the family offers, without technical details. The data come from Google's and Ollama's official pages consulted on 7 October 2026.

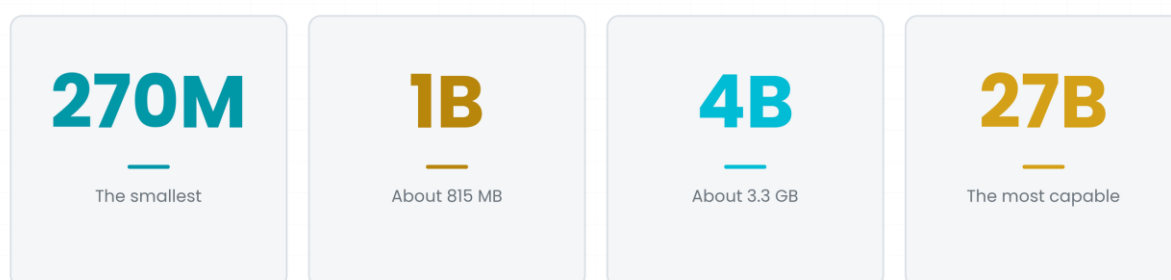
3.1 Gemma 4 (April 2026)

- Released on **2 April 2026**, under the **Apache 2.0** licence (very permissive, also for commercial use).
- Available in several sizes: **E2B** and **E4B** (the smallest), **26B** mixture-of-experts and **31B** dense. In June 2026 a **12-billion**-parameter version was added.
- It is **multimodal**: it also understands images and video, and the small sizes also audio.
- Context up to **256 thousand tokens** in the larger sizes (the «memory» of a conversation).

3.2 Gemma 3 (the previous generation)

- Very small sizes: **270 million**, **1 billion**, **4**, **12** and **27 billion** parameters.
- Supports **over 140 languages**.
- From the 4-billion size up it also understands **images**.
- The smallest versions are meant exactly for computers with few resources: **815 MB** the 1-billion model, **3.3 GB** the 4-billion one.

Gemma 3: from small to large



The sizes of Gemma 3

WHAT ARE THE BILLIONS?

The number (1B, 4B, 27B...) counts the «parameters», the model's internal knobs. The more there are, the more capable the model, but the bigger, slower and hungrier for memory.

4. How it is launched, in broad terms

The simplest way to run an open model is a free program called **Ollama**. It is a bit like a music player: you install the player once, then you choose the tracks, that is the models.

- You download Ollama from the official site and install it.
- You choose a model from its online library, looking at **size and requirements**.
- You launch it with a command, for example **ollama run** followed by the model's name, and a chat opens in the terminal or in the app window.

I have Ollama on the computer I use for this course, but with code models, not with Gemma, and **I did not try Gemma for the lesson**: here I only tell you what is there and what to expect. If you want to try, start from the smallest model you find, and see how it answers on your computer.

DOWNLOAD ONLY FROM OFFICIAL SITES

As with the Gemini app, download Ollama and the models only from official sites (ollama.com). Beware of sites offering «free Gemma» with odd programs: they are scam attempts.

5. When it makes sense (and when not)

Situation	Better
Very confidential texts that cannot leave the computer	Gemma locally
Working offline, travelling or without a stable connection	Gemma locally
I want the most capable model, web search and connected apps	Gemini in the cloud
I have an old or low-powered computer	Gemini in the cloud
I want to experiment, learn how a model works	Gemma locally , starting from a small size

How to choose between Gemini and Gemma.

For most people, **Gemini in the cloud remains the most convenient and most capable choice**. Gemma is fascinating for those with privacy needs, for those who study and for those who want to build something of their own.

Hands-on exercise

Fifteen minutes, without downloading anything: just read and decide.

1. Open the official Gemma page and the Ollama library page and read the available sizes with their dimensions.
2. Check how much memory and which graphics card your computer has (in Windows: Task Manager, Performance tab).
3. Write three things you would gladly do locally and three you prefer to do with Gemini in the cloud.
4. Decide: do I try Gemma, yes or no? If yes, which size and when?
5. If you decide to try, download only the smallest model first, and only from official sites.

WATCH OUT

Can you tell the difference between open weights and a model in the cloud?

Did you check the free space before deciding?

Questions that come up in class

Is Gemma free?

Yes, Gemma models can be downloaded and used, and the Gemma 4 licence is Apache 2.0. The real cost is your hardware and electricity.

Is Gemma the same as Gemini?

No. They are different Google model families. Gemma is open and downloadable; Gemini is closed and used from the cloud.

Do I need a powerful graphics card?

For small models less is enough, but in general the bigger the model, the more memory and speed you need. Start from the small sizes.

Does it work without internet?

Yes, after downloading the model. To download it you need a connection.

Can I connect it to Gmail or Drive?

Not directly like Gemini. Locally you have the model, not Google's connected tools.

Glossary

Term	What it means
Gemma	Google's family of open models, which can be downloaded and run locally.
Locally	On your computer, without sending texts to a cloud service.
Ollama	A free program to download and run open models on your own computer.
Open weights	The model's internal numbers, made public so they can be downloaded and used.
Parameters	A model's internal knobs: their number indicates the size.

Further reading

Google, Gemma 4: Byte for byte, the most capable open models (blog) - announcement of 2 April 2026

Google AI for Developers, Run Gemma with Ollama - official page

Ollama, library of the gemma3 and gemma4 models - pages consulted on 7 October 2026 for sizes and file dimensions

WATCH OUT

Data checked on 7 October 2026 on the official pages. In this lesson no Gemma model was downloaded or tried: the features are described by the sources. File sizes and available versions change often.