

Gemma: l'intelligenza artificiale di Google sul tuo computer

Un assaggio dei modelli aperti, che girano in locale, senza abbonamento e senza internet

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



Indice

Come usare questa dispensa.....	3
1. Gemini o Gemma?.....	3
2. Perché farlo girare in locale.....	4
3. Le funzionalità delle famiglie Gemma.....	5
3.1 Gemma 4 (aprile 2026)	5
3.2 Gemma 3 (la generazione precedente)	5
4. Come si lancia, a grandi linee	6
5. Quando ha senso (e quando no).....	6
Esercitazione pratica.....	7
Le domande che arrivano in aula.....	8
Glossario	8
Per approfondire	9

Come usare questa dispensa

Questa dispensa raccoglie, in forma scritta e ordinata, tutto quello che vediamo nella lezione. Non serve studiarla a memoria: serve a ritrovare le cose con calma, dopo, quando la lezione è finita e resta una domanda in sospeso.

OBBIETTIVO

Capire la differenza tra Gemini e Gemma, che cosa significa far girare un modello in locale e quando ha senso farlo, senza entrare nei dettagli dell'installazione.

Che cosa saprai fare alla fine

- Spiegare la differenza tra Gemini (nel cloud) e Gemma (modelli aperti).
- Elencare vantaggi e limiti di un modello che gira sul proprio computer.
- Riconoscere le caratteristiche principali delle famiglie Gemma 3 e Gemma 4.
- Sapere che cos'è Ollama e come si lancia un modello, a grandi linee.

1. Gemini o Gemma?

Gemini è un ristorante: ordini, ti servono, paghi (o no). Gemma è la ricetta: te la porti a casa e cucini tu.

Finora abbiamo parlato di **Gemini**, che vive nei computer di Google: tu scrivi, la richiesta parte via internet, la risposta torna. **Gemma** è un'altra famiglia di modelli di Google, con una differenza fondamentale: è a **pesi aperti**. I «pesi» sono il cuore numerico del modello, e quando sono aperti puoi **scaricarli** e farli girare sul tuo computer.

IN PAROLE SEMPLICI

Gemini lo usi dal cloud. Gemma lo porti in casa. Sono parenti, non la stessa cosa: Gemma non è «Gemini in versione locale».

Il ristorante e la ricetta

Gemini

- Nel cloud di Google
- Strumenti e app collegate
- Piani e limiti d'uso

vs

Gemma

- Sul tuo computer
- Pesi aperti, scaricabili
- Nessun abbonamento

Gemini e Gemma a confronto

2. Perché farlo girare in locale

Far girare un modello sul proprio computer si dice **in locale**. Ha vantaggi veri e limiti altrettanto veri.

Vantaggi	Limiti
Privacy: i testi non lasciano il computer.	Hardware: servono memoria e, meglio, una buona scheda video.
Funziona offline , dopo il download.	Qualità: un modello piccolo è meno capace di uno grande nel cloud.
Nessun abbonamento né limiti d'uso del piano.	Niente strumenti collegati: niente Gmail, Drive, ricerca web come in Gemini.
Libertà: puoi provarlo, adattarlo, usarlo in programmi tuoi.	Manutenzione: aggiornare, scaricare e gestire i file tocca a te.

In locale: che cosa guadagni e che cosa perdi.

IL PESO DEI FILE

I modelli sono file grandi: da qualche centinaio di megabyte a decine di gigabyte, secondo le dimensioni. Prima di scaricare, controlla lo spazio libero e la dimensione indicata sulla pagina ufficiale del modello.

3. Le funzionalità delle famiglie Gemma

Ti faccio un assaggio di che cosa offre la famiglia, senza entrare nei dettagli tecnici. I dati vengono dalle pagine ufficiali di Google e di Ollama consultate il 7 ottobre 2026.

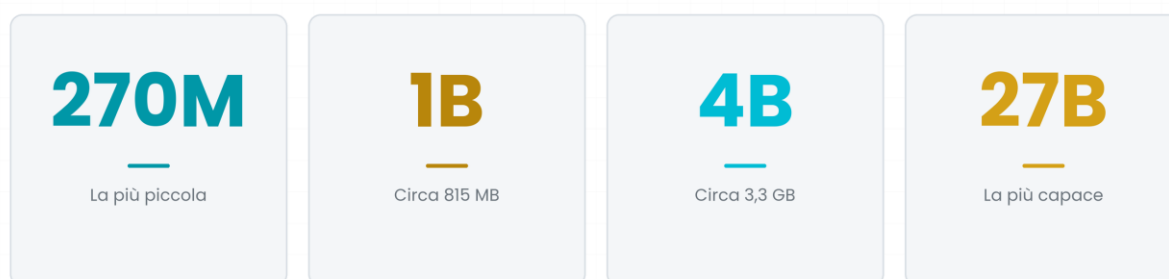
3.1 Gemma 4 (aprile 2026)

- Rilasciata il **2 aprile 2026**, con licenza **Apache 2.0** (molto libera, anche per usi commerciali).
- Disponibile in più taglie: **E2B** e **E4B** (le più piccole), **26B** a miscela di esperti e **31B** densa. A giugno 2026 è stata aggiunta una versione da **12 miliardi** di parametri.
- È **multimodale**: capisce anche immagini e video, e le taglie piccole anche l'audio.
- Contesto fino a **256 mila token** nelle taglie più grandi (la «memoria» di una conversazione).

3.2 Gemma 3 (la generazione precedente)

- Taglie molto piccole: **270 milioni**, **1 miliardo**, **4**, **12** e **27 miliardi** di parametri.
- Supporta **oltre 140 lingue**.
- Dalla taglia da 4 miliardi in su capisce anche le **immagini**.
- Le versioni più piccole sono pensate proprio per computer con poche risorse: **815 MB** il modello da 1 miliardo, **3,3 GB** quello da 4 miliardi.

Gemma 3: dal piccolo al grande



Le taglie di Gemma 3

CHE COSA SONO I MILIARDI?

Il numero (1B, 4B, 27B...) conta i «parametri», le manopole interne del modello. Più sono, più il modello è capace, ma più grande, lento e affamato di memoria.

4. Come si lancia, a grandi linee

Il modo più semplice per far girare un modello aperto è un programma gratuito che si chiama **Ollama**. È un po' come un lettore di musica: installi il lettore una volta, poi scegli i brani, cioè i modelli.

- Scarichi Ollama dal sito ufficiale e lo installi.
- Scegli un modello dalla sua libreria online, guardando **dimensione e requisiti**.
- Lo lanci con un comando, per esempio **ollama run** seguito dal nome del modello, e si apre una chat nel terminale oppure nella finestra dell'app.

Io ho Ollama sul computer che uso per questo corso, ma con modelli per il codice, non con Gemma, e **non ho provato Gemma per la lezione**: qui ti dico solo che cosa c'è e che cosa aspettarti. Se vuoi provare, parti dal modello più piccolo che trovi, e vedi come risponde sul tuo computer.

SCARICA SOLO DAI SITI UFFICIALI

Come per l'app di Gemini, scarica Ollama e i modelli solo dai siti ufficiali (ollama.com). Occhio ai siti che offrono «Gemma gratis» con programmi strani: sono tentativi di truffa.

5. Quando ha senso (e quando no)

Situazione	Meglio
Testi molto riservati che non possono uscire dal computer	Gemma in locale
Lavoro offline, in viaggio o senza connessione stabile	Gemma in locale
Voglio il modello più capace, ricerca web e app collegate	Gemini nel cloud
Ho un computer vecchio o poco potente	Gemini nel cloud
Voglio sperimentare, imparare come funziona un modello	Gemma in locale , partendo da una taglia piccola

Come scegliere tra Gemini e Gemma.

Per la maggior parte delle persone, **Gemini nel cloud resta la scelta più comoda e più capace**. Gemma è affascinante per chi ha esigenze di privacy, per chi studia e per chi vuole costruire qualcosa di proprio.

Esercitazione pratica

Quindici minuti, senza scaricare nulla: solo leggere e decidere.

1. Apri la pagina ufficiale di Gemma e quella della libreria di Ollama e leggi le taglie disponibili con le dimensioni.
2. Controlla quanta memoria e che scheda video ha il tuo computer (in Windows: Gestione attività, scheda Prestazioni).
3. Scrivi tre cose che faresti volentieri in locale e tre che preferisci fare con Gemini nel cloud.
4. Decidi: provo Gemma, sì o no? Se sì, quale taglia e quando?
5. Se decidi di provare, scarica prima solo il modello più piccolo, e solo dai siti ufficiali.

ATTENZIONE

Sai dire la differenza tra pesi aperti e modello nel cloud?

Hai controllato lo spazio libero prima di decidere?

Le domande che arrivano in aula

Gemma è gratis?

Sì, i modelli Gemma si possono scaricare e usare, e la licenza di Gemma 4 è Apache 2.0. Il costo reale è il tuo hardware e la corrente.

Gemma è la stessa cosa di Gemini?

No. Sono famiglie di modelli diverse di Google. Gemma è aperta e si scarica; Gemini è chiusa e si usa dal cloud.

Serve una scheda video potente?

Per i modelli piccoli basta meno, ma in generale più il modello è grande, più serve memoria e velocità. Parti dalle taglie piccole.

Funziona senza internet?

Sì, dopo aver scaricato il modello. Per scaricarlo la connessione serve.

Posso collegarla a Gmail o a Drive?

Non direttamente come Gemini. In locale hai il modello, non gli strumenti collegati di Google.

Glossario

Termine	Che cosa significa
Gemma	La famiglia di modelli aperti di Google, che si possono scaricare e far girare in locale.
In locale	Sul tuo computer, senza mandare i testi a un servizio nel cloud.
Ollama	Un programma gratuito per scaricare e far girare modelli aperti sul proprio computer.
Parametri	Le manopole interne di un modello: il loro numero ne indica la taglia.
Pesi aperti	I numeri interni del modello, resi pubblici in modo da poterli scaricare e usare.

Per approfondire

Google, Gemma 4: Byte for byte, the most capable open models (blog) - annuncio del 2 aprile 2026

Google AI for Developers, Run Gemma with Ollama - pagina ufficiale

Ollama, libreria dei modelli gemma3 e gemma4 - pagine consultate il 7 ottobre 2026 per taglie e dimensioni dei file

ATTENZIONE

Dati verificati il 7 ottobre 2026 sulle pagine ufficiali. In questa lezione non è stato scaricato né provato alcun modello Gemma: le funzionalità sono descritte dalle fonti. Dimensioni dei file e versioni disponibili cambiano spesso.