Local AI Challenges

LLM locali, privacy, RAG e agenti su Linux: cosa possiamo già fare con hardware "normale" .



Samuele Pacioni
Software Engineer
git @SamPac-Dev

FermoLUG
linuxfm.org

LinuxDay
Italia2026

LICENZA CC BY-NC-SA 4.0

Che cos'è un LLM?

Un Large Language Model è un modello linguistico addestrato su grandi quantità di testo e apprende schemi del linguaggio.



Testi
Addestramento su grandi quantità di testo.

Modello
Apprende schemi del linguaggio.

Elaborazione
Elabora il contesto della domanda.

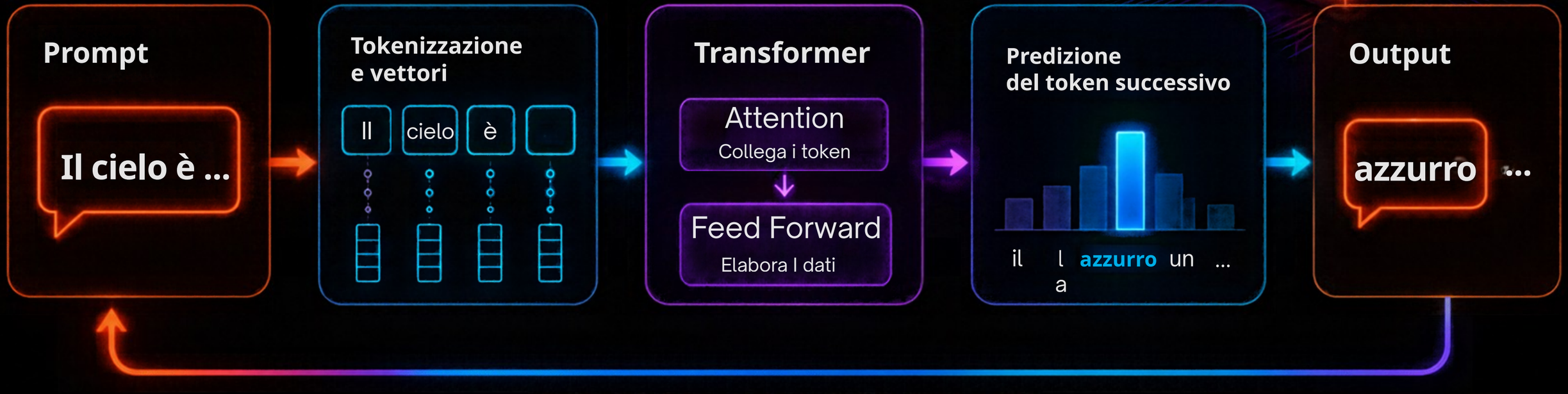
Un LLM è in grado di **comprendere il contesto** di una richiesta (*prompt*) e generare risposte coerenti attraverso la previsione delle parole successive.

ATTENZIONE !
può generare imprecisioni (**allucinazioni**)

Come funziona un LLM?

Da un prompt in ingresso, il testo viene convertito in token e vettori, elaborato nel Transformer tramite attention e feed forward, viene previsto probabilisticamente il token successivo e si ottiene l'outpuput.
Il ciclo si ripete per generare i token successivi





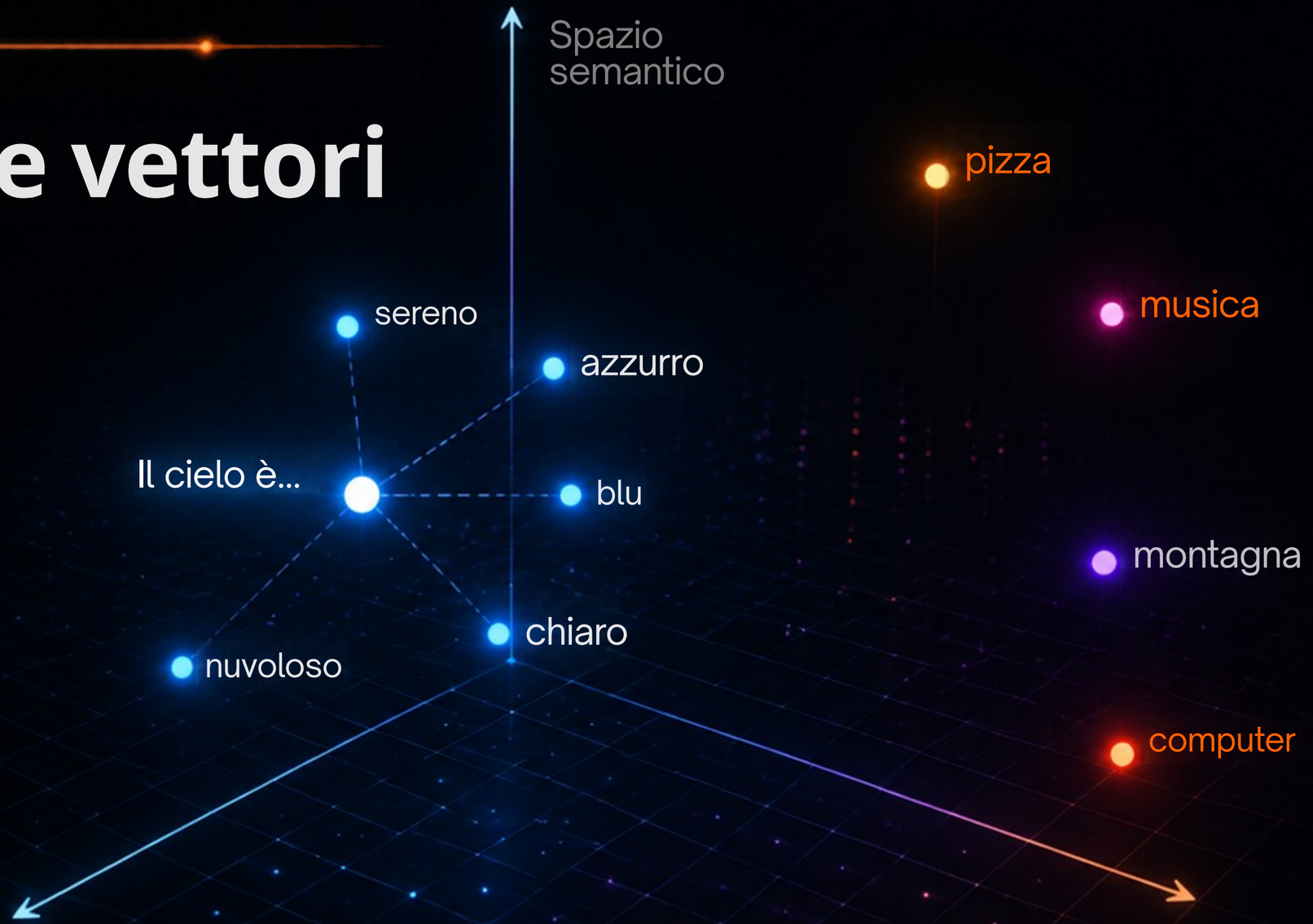
Il ciclo si ripete per generare i token successivi.

Le parole come vettori

Il modello rappresenta le parole come vettori in uno spazio semantico: termini simili sono vicini, termini non correlati sono lontani.

Nel completamento «Il cielo è...» le parole come «azzurro», «blu», «sereno», «nuvoloso» e «chiaro» sono vicine alla frase, mentre termini non correlati sono distanti.

Il modello usa le distanze semantiche per stimare la continuazione più probabile.



Dai grandi LLM agli Small Language Models

PASSATO

Richiedeva datacenter, GPU da server, raffreddamento e competenze specialistiche.

dal datacenter al computer che hai già

OGGI

Un laptop può eseguire modelli quantizzati e software scaricabile. I requisiti variano secondo modello e configurazione.



GRANDI LLM
Elevata capacità · Elevate risorse

SMALL LANGUAGE MODELS (SLM)
Modelli compatti e ottimizzati · 1B–8B parametri

LOCAL AI
Laptop · Smartphone · Mini PC



Più grande non significa sempre più adatto.

Open weight non significa open source

Pesi disponibili

I pesi possono essere scaricabili e utilizzabili.



Dati di training non disponibili

I dati e la ricetta di addestramento possono restare proprietari.





Ti danno la torta, non gli ingredienti né la ricetta.

Dall’open weight al modello scaricabile



Il laboratorio/produttore raccoglie grandi quantità di testi, codice e immagini.

Il training modifica i parametri del modello.

I pesi pubblicati permettono di scaricare ed eseguire il modello.

Fonti per Modelli Open Weight

Dove trovare modelli affidabili e aggiornati:

- **Ollama Library** (ollama.com/library) — catalogo curato, download diretto con un comando, aggiornamenti automatici
- **Hugging Face** (huggingface.co) — il più grande repository OSS, migliaia di modelli, schede dettagliate con benchmark

- Criteri per la scelta del modello giusto:
- — Dimensione compatible con la RAM/VRAM disponibile
- — Quantizzazione adeguata (4/8/16 bit) al caso d'uso
- — Data di rilascio e attività della community
- — Licenza d'uso (MIT, Apache 2.0, licenze custom)
- — Valutazioni su benchmark specifici per il task



RAM **Dimensione** compatibile con la RAM/VRAM

Quantizzazione adeguata (4/8/16 bit)

Data di rilascio e attività della community

Licenza d'uso (MIT, Apache 2.0, licenze custom)

Valutazioni su benchmark specifici per il task

Perché conviene possedere il modello



Nessun costo mensile per abbonamenti cloud AI



Dati sotto il controllo dell’utente



Possibilità di lavorare offline



Personalizzazione tramite modifiche o fine-tuning

I modelli cloud offrono potenza immediata, ma richiedono connessione, costi variabili e condivisione dei dati.
I modelli locali e open source restituiscono controllo, privacy e autonomia all'utente.

Hardware reale: CPU, GPU 4 GB, eGPU 16 GB



CPU e RAM

GPU 4 GB

eGPU 16 GB

Ollama: Runtime leggero per modelli locali

Ollama semplifica l'esecuzione di LLM open source in locale: un unico strumento per installare, scaricare, gestire e avviare modelli con pochi comandi.



Installazione e gestione

Setup immediato su Linux, Windows e macOS. Download, cache e avvio dei modelli con un solo comando: ollama run.

API REST compatibile OpenAI

Espone un'API locale sulla porta 11434, compatibile con le specifiche OpenAI. Integrabile con qualsiasi client o tool che supporti OpenAI.

CPU/GPU

Supporto nativo CPU e GPU (NVIDIA/AMD). Gestisce quantizzazioni GPU: 4, 8 e 16 bit per adattarsi all'hardware disponibile.

AnythingLLM

Ollama è il motore: gestisce i modelli, l'inferenza e le API locali. **AnythingLLM** è lo strato superiore: interfaccia, workspace, strumenti e contesto documentale. Insieme trasformano un modello grezzo in un assistente realmente usabile. Open source e Installazione Self-Hosted.



Ollama
Motore locale: pesi, inferenza, API REST compatibile OpenAI

→

AnythingLLM
interfaccia + workspace + strumenti + gestione documenti

→

Assistente Completo
Il modello diventa usabile, contestuale e personalizzato

Architettura Ollama + AnythingLLM



User AnythingLLM Ollama Local Model GPU





Documents and Tools

Il Workspace

Un workspace è un contenitore isolato per progetti e contesti specifici.
Esempi di workspace:

- Documenti FermoLUG — verbali, presentazioni, comunicati
- Manuali tecnici — guide operative e istruzioni
- Procedure interne — processi aziendali e flussi di lavoro
- Listini e cataloghi — prodotti, prezzi, offerte
- Knowledge base — FAQ, note, archivo storico

Ogni workspace ha il proprio contesto personalizzato: i documenti caricati restano separati dagli altri workspace, garantendo risposte precise e pertinenti al progetto.



Ma una chat non basta

👤 Utente: "Qual è la procedura interna per l'approvazione dei fornitori?"
🌹 LLM: "Non lo so. Non ho accesso ai tuoi documenti interni."
Il modello linguistico conosce il mondo, ma non conosce la tua azienda. Senza i tuoi documenti, non può rispondere a domande specifiche sul tuo contesto.
Serve un modo per portare la conoscenza interna dentro il modello.



RAG spiegato semplice

Retrieve & Generate: prima cerca, poi risponde. Il modello non memorizza tutto, recupera ciò che serve al momento giusto.

Flusso: domanda → ricerca nei documenti → estrazione dei pezzi rilevanti → input al modello → risposta contestuale.



Documenti

Caricamento di file e documenti aziendali.

Retrieve/Embedding

Ricerca e indicizzazione per trovare i contenuti più pertinenti.

Frammenti rilevanti

Selezione delle parti più utili per rispondere alla domanda.

LLM

Il modello elabora il contesto e genera la risposta.

Risposta

Risposta precisa e contestualizzata al tuo questo.

Da Chatbot ad Agent



Un LLM tradizionale risponde solo con ciò che sa. Un agent può scegliere e usare strumenti esterni: filesystem, database SQL, web, documenti. Prima rispondeva, ora può anche agire.



LLM riceve la domanda e sceglie il tool più adatto



Tool essegue l'azione: filesystem, SQL, web, documenti



LLM elabora il risultato e fornisca la risposta finale

Oltre il Testo

Local AI non è solo una chat: è un laboratorio completo. Analizza immagini, trascrive voci, legge pagine web e crea documenti strutturati.





Visione e Immagini

Analisi contenuto fotografico, ricerca immagini per soggetto, generazione e modifica visuale con modelli multimodali.



Voce e Web

Trascrizione audio locale con Whisper, lettura e sintesi di pagine web, estrazione dati da fonti online.



Documenti e Output

Creazione report, sintesi automatiche, esportazione strutturata. Locali AI genera output pronti all'uso, non solo risposte.

Le demo del laboratorio

Dal modello locale agli agenti: un percorso pratico, passo dopo passo

MODELLI LOCALI E INTERFACCIA

Demo 0
LLM locale con Ollama

Demo 1
LLM locale con AnythingLLM

Workspace: INTRO
Modello: MinVerva:7B

RAG E AGENT WEB

Demo 1.2
AI Agent Web Simple
Workspace: WEBSEARCH2
Modello: gemma4:e4b

Demo 2
RAG: catalogo e-bike
Workspace: RAG_E-BIKE
Modello: Ministral:14B

Demo 2.1
RAG_URL
Workspace: RAG_URL
Modello: Qwen3:8B

ANALISI DATI, VISIONE E DATA-FLOW

Demo 3
Analisi delle vendite 2026
Workspace: SQLCONNECT
Modello: Qwen3.5:9B

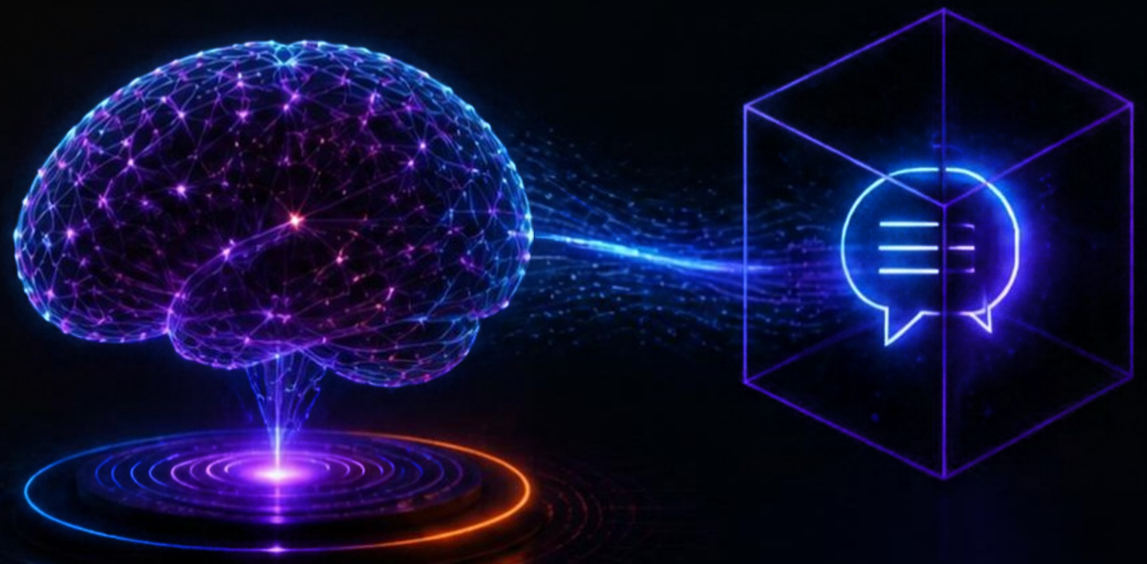
Demo 4
Vision AI
Workspace: IMAGE ELAB
Modello: Minicpm-v4.5

Demo 5
AI Agent Web
Modello: gemma4:e4b

Demo 6
Data-Flow
Workspace: DATAFlow
Modello: Ministral:14B

Cosa ho Imparato

Cinque insight chiave emersi durante la preparazione e le demo di questo laboratorio su LLM locali, agent e RAG con strumenti open source.



Più grande non sempre significa migliore

Un modello piccolo e ben quantizzato spesso supera un modello enorme in termini di velocità e usabilità pratica. I documento contano più del modello: con RAG, la qualità dei dati è tutto.

Gli agent vanno testati e limitati

Dare agli agent solo i permessi minimi necessari. Le demo richiedono test ripetuti: un prompt che funziona oggi potrebbe non funzionare domani senza un ambiente stabile.

Linux è un ottimo laboratorio AI

Linux offre controllo totale su runtime, modelli e strumenti. È la piattaforma ideale per sperimentare con LLM locali, OLLama, AnythingLLM e agent personalizzati.

Parametri, Precisione e Esempi di Modelli

I modelli LLM variano per dimensione e precisione dei pesi: Meno bit = meno memoria, ma anche meno precisione. La quantizzazione permette di eseguire modelli grandi su hardware consumer.

Il Motore di Inferenza

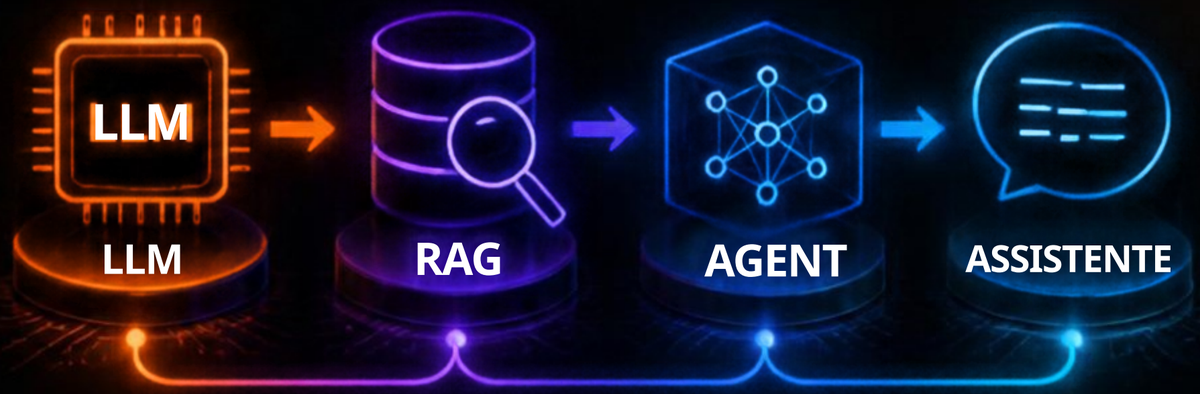
Il motore di inferenza gestisce l'intero ciclo di elaborazione: carica i pesi del modello in memoria, e restituisce la risposta token per token. Latenza, throughput e utilizzo della RAM dipendono direttamente dalla sua efficienza.

Conclusioni

Che AI posso usare? Dove gira? Con quali dati? Sotto il controllo di chi?

LLM locale → RAG → Agent → Assistente sui tuoi dati.

La tecnologia è disponibile, open source e funziona oggi.



Sotto il controllo di chi?
Tu i tuoi dati restano dove decidi tu.

Con quali dati?
I tuoi documenti, i tuoi sistemi, le tue regole.

Dove gira?
Sul tuo hardware, in locale o cloud privato.

Tecnologia oggi
Open source, matura, pronta da usare.

Domande ?

Grazie a FermoLUG e a tutti i partecipanti !

Samuele Pacioni
Software Engineer
git *@SamPac-Dev*



Risorse utili per iniziare con l’AI locale

Download, cataloghi, requisiti hardware e guide pratiche

SOFTWARE · MODELLI · GUIDE

ANYTHINGLLM · Download
App desktop: chat, workspace, RAG e agent
anythingllm.com/download

OLLAMA · Installazione
Runtime locale per Linux, Windows e macOS
ollama.com/download

OLLAMA · Libreria modelli
Catalogo LLM e comandi per scaricarli
ollama.com/library

HUGGING FACE · Modelli
Schede, licenze, varianti e repository
huggingface.co/models

ANYTHINGLLM · Documentazione
Tutorial su RAG, agent e configurazione
docs.anythingllm.com

VERIFICA HARDWARE · PRESTAZIONI

llmrun
GPU, VRAM e stime di velocità (tok/s)
llmrun.dev

Run This LLM
Quali modelli girano sul mio hardware?
runthisllm.com

Can I Use LLM?
CPU, RAM e quantizzazione consigliata
caniusellm.com

Can I Run AI Locally?
GPU e memoria: modelli compatibili
canirun.ai

OLLAMA-MONITOR
Monitor Ollama e Risorse GPU by Samuele Pacioni
https://github.com/SamPac-Dev/Ollama-Monitor

Appendice 1 | LLM: le parole chiave

Dal prompt ai token, dai parametri alla risposta

TOKEN

Unità di testo: può essere una parola o una sua parte.

PARAMETRI

7B = circa 7 miliardi di parametri, non 7 GB di memoria.

PROMPT

Domanda o istruzioni fornite al modello.

CONTEXT WINDOW

Massimo contesto gestibile, misurato in token.

TEMPERATURA

Regola la variabilità delle risposte, non la loro accuratezza.

QUANTIZZAZIONE

Riduce la precisione dei pesi e la memoria richiesta (es. Q4).

INFERENZA

Il modello già addestrato elabora il prompt e genera l'output.

EMBEDDING

Vettore numerico utile per cercare contenuti simili nel RAG.

Appendice 2 | Modelli LLM open-weight fino a 14B

Famiglie diffuse per AI locale • produttore, Paese, modalità e licenza

MODELLO (PARAMETRI)	PRODUTTORE	PAESE	INPUT → OUTPUT	LICENZA
Gemma 4 · E2B / E4B / 12B	Google DeepMind	USA	Testo, immagini, audio → testo	Apache 2.0
Llama 3.1 / 3.2 · 1–11B	Meta	USA	Testo → testo; 11B: immagini	Llama Community
Ministral 3 · 14B	Mistral AI	Francia	Testo / immagini → testo	Apache 2.0
Phi-4 · 14B	Microsoft	USA	Testo → testo	MIT
Qwen2.5 · 7B / 14B	Alibaba Cloud	Cina	Testo → testo	Apache 2.0
Qwen3.5 · 0,8–9B	Alibaba Cloud	Cina	Testo / immagini → testo	Apache 2.0
Velvet · 14B	Almawave	Italia	Testo → testo	Apache 2.0

Le modalità dipendono dalla variante. Open-weight ≠ necessariamente open-source; verificare sempre la licenza specifica.

Appendice 3 - Dalla domanda al risultato

Un prompt efficace chiarisce ruolo, compito, regole, output e controlli

01

RUOLO

Un buon prompt è un insieme di istruzioni.

Agisci come un analista visivo

02

COMPITO

Specificare obiettivo e risultato atteso aiuta l'AI a svolgere il compito.

Analizza tutte le immagini

03

REGOLE

Definire limiti riduce il rischio di informazioni inventate.

Non inventare brand illeggibili

04

OUTPUT

Chiedere un formato preciso rende il risultato più facile da verificare.

Restituisci una tabella

05

CONTROLLI

Aggiungere controlli finali migliora coerenza e affidabilità.

Conta i brand distinti

Più istruzioni chiare → risultati più coerenti, utili e verificabili.

Appendice 4 - Prompt in azione

Esempio: l’AI analizza le foto delle bevande

Le istruzioni definiscono il **ruolo, il compito** e le **categorie** da usare.

Il vincolo **“Non inventare brand illeggibili”** riduce le supposizioni.

L’output richiesto è **una riga per immagine**, con categoria, brand e descrizione.

Il conteggio finale riassume i **brand distinti leggibili**.

PROMPT

Sei un assistente che analizza foto di bevande.

Il tuo compito è identificare la categoria, il brand e fornire una breve descrizione.

Usa queste categorie: VINO, BIRRA, BIBITA.

Non inventare brand illeggibili.

Rispondi con una riga per immagine, con categoria, brand e descrizione.

Alla fine, fornisci il conteggio dei brand distinti leggibili.

RISULTATO

N°	Categoria	Brand	Descrizione
1	VINO	Collina Rossa	Bottiglia su tavolo
2	BIRRA	Luppolo 21	Lattina dorata
3	BIBITA	Non leggibile	Bottiglia in plastica

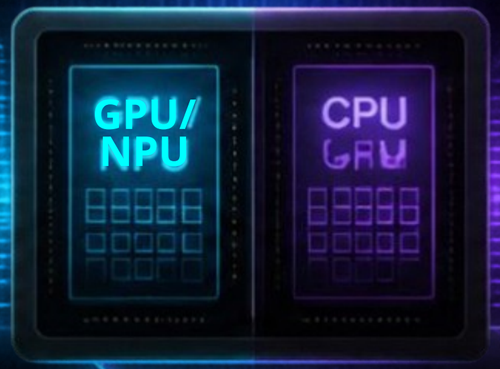
Totale brand distinti leggibili: 2

Un prompt dettagliato riduce l’ambiguità, ma non elimina del tutto gli errori dell’AI.

Appendice 5
VRAM e memoria unificata: scegliere l'hardware

La memoria disponibile determina quali modelli locali puoi eseguire.







	1	2	3	4	5
CAPACITÀ (Memoria)	4 GB	8 GB	12-16 GB	24-32 GB	64-128 GB
MODELLO (ESEMPI)	Qwen 2.5 GB	Mistral 7B	Qwen 2.5 14B	Qwen 2.5 32B /	Mixtral 8x22B
MEMORIA STIMATA (Q4_K_M)	≈ 2,2 GB	≈ 4,7 GB	≈ 9-15 GB	≈ 20-28 GB	≈ 43-90 GB

VRAM dedicata
GPU consumer (es. NVIDIA/AMD)

1 → 2 → 3 → 4 → 5

Memoria unificata
Apple Silicon (M-series) e
AMID Ryzhen AI Max / Strix Halo

1 → 2 → 3 → 4 → 5

Appendice 6
Il collo di bottiglia: la larghezza di banda

La velocità di generazione dei token dipende in larga parte dalla larghezza di banda della memoria.

VRAM dedicata → memoria unificata



VRAM dedicata
Ampia banda, capacità limitata.

Memoria unificata
Capacità elevata, banda condivisa.

Confronto teorico della larghezza di banda (GB/s)





Desktop con GPU dedicata

Velocità elevata e modelli fino a circa 14B-32B.

Mini-PC o laptop AMD Ryzen AI Max

Grande capacità di memoria unificata con compromesso economico.

Mac Studio / Apple Silicon

Massima capacità di memoria unificata per modelli più grandi.