iaLocale

Guida all'IA Locale per Aziende

Hardware, modelli, strumenti e primi passi: tutto quello che serve per portare l'IA nel tuo lavoro — senza cloud e senza condividere i dati. Con prezzi reali e comandi pronti da copiare.

iaLocale · guida gratuita · 1ª edizione, settembre 2026

ialocale-it.apps.tommasoconcato.com/guida

CONTENUTI

Indice

1. Cos'è l'IA locale (e cosa non è)
2. Le tre domande da fare prima di comprare
3. Hardware: come leggere le specifiche
4. Modelli e quantizzazione: come scegliere
5. Strumenti: Ollama, Open WebUI, vLLM
6. Primi passi: 10 minuti, due comandi
7. I limiti: quando l'IA locale non basta
8. Dove andare avanti

Chi è il lettore di questa guida: chi ha un'attività, un team o un progetto e vuole capire se l'IA locale gli conviene — senza un budget IT dedicato e senza termini che nessuno spiega. Se invece cerchi l'installazione passo-passo dei singoli strumenti, i link in fondo alla guida puntano al materiale completo.

1

Cos'è l'IA locale (e cosa non è)

L'IA locale è un modello linguistico che gira sul **tuo** computer o sul tuo server — non su quello di qualcun altro. Invece di inviare la tua domanda a un cloud, il calcolo avviene sulla tua GPU (o sulla memoria del tuo Mac), e la risposta torna a te senza che nessun byte lasci la stanza.

Un'immagine: la differenza tra affittare e comprare

Usare un servizio cloud è come **affittare**: paghi a ogni utilizzo, il prezzo può cambiare, e il contratto dice che i tuoi dati passano per le mani del proprietario. L'IA locale è **comprare**: il costo è l'hardware (una tantum) e i dati restano tuoi. Chi ha budget illimitati può preferire l'affitto; chi ha costi ricorrenti o dati sensibili, quasi sempre il contrario.

I tre vantaggi concreti

- **Controllo dei dati.** Clienti, fatture, contratti: nulla lascia la tua infrastruttura. Non c'è log del fornitore su cui ragionare, non c'è dati che finiscono in training altrui.
- **Costo prevedibile.** Dopo l'acquisto dell'hardware, l'inferenza costa la corrente. Un modello che serve 10 persone al giorno su un'GPU da 800 euro costa, in pratica, zero a richiesta.
- **Disponibilità.** Funziona quando internet cade, in ufficio, in stabilimento, nel cantiere. Nessun limite di richieste, nessun picco di prezzo, nessuna dipendenza da un vendor che chiude il servizio.

> **La parte onesta:** i modelli locali non eguagliano i frontier model cloud per i task più difficili (lunghi ragionamenti, compiti scientifici complessi). E costano di più in partenza. La domanda giusta non è "cos'è meglio?", ma "per il mio uso specifico, cosa mi costa di meno e cosa mi conviene di più?". È quello che rispondiamo insieme nelle prossime pagine.

2

Le tre domande da fare prima di comprare

Prima di spendere un euro, rispondi a tre domande. Sono le stesse che usa il pianificatore di setup di iaLocale, e determinano il 90% della scelta.

Domanda 1 — Che compito deve fare?

Riassunti di documenti, bozze di email, ricerca in una base di documenti, un agente che chiama strumenti: sono compiti con esigenze diverse. I riassiunti e il chat richiedono poco; un agente che ragiona in più passi e un contesto lungo richiedono molto di più. Il task definisce la *dimensione del modello*.

Domanda 2 — Quanto deve essere veloce?

Una risposta in 3 secondi e una in 15 secondi sono esperienze diverse per chi lavora. La velocità dipende dalla *banda della memoria* dell'hardware, non dalla potenza di calcolo — è il punto che stupisce tutti, ed è spiegato nella sezione hardware.

Domanda 3 — Quanti utenti, in parallelo?

Un modello serve un utente alla volta. Se sono cinque persone a scrivere nello stesso momento, o serve il contesto lungo di tutta la conversazione, serve più memoria — o un'infrastruttura che gestisce la coda.

> **Il metodo:** prendi un caso d'uso reale (il riassunto del verbale, la bozza dell'offerta, la ricerca nei manuali di prodotto) e definiscilo in numeri: quanto testo entra, quante parole escono, in quanto tempo deve arrivare, quante persone lo usano. Con quei numeri, le sezioni 3 e 4 diventano una scelta, non un'ipotesi.

Un esempio concreto

Un'agenzia di 8 persone che vuole: bozze di email e preventivi, ricerca nei contratti, un assistente in italiano. Profilo: modello 8–14B in quantizzazione Q4, 16 GB di VRAM, banda 800+ GB/s. Esito: una GPU RTX 4070 Ti Super (16 GB, ~800 euro) copre il caso con margine. Lo stesso budget su un cloud sarebbe un abbonamento ricorrente che, a questi volumi, costa più dell'hardware nel primo anno.

3

Hardware: come leggere le specifiche

Per l'IA locale contano **due numeri** dell'hardware: la VRAM (quanta memoria ha la scheda) e la banda (quanto velocemente la memoria si legge). Tutto il resto è secondario.

VRAM: decide QUANTO modello entra

La regola pratica: i pesi di un modello occupano circa `parametri × bit / 8` byte. Un modello da 8 miliardi di parametri in Q4_K_M pesa ~4,8 GB di pesi, più la memoria di lavoro (il contesto, la KV cache). Per lavorarci con margine: **modello 8B → 12–16 GB; modello 14B → 16–24 GB; modello 30B → 24–32 GB**.

Banda: decide QUANTO è veloce

Le parole escono a una velocità proporzionale alla banda della memoria, divisa per i byte del modello. Un'RTX 3060 (360 GB/s) su un 8B Q4 fa ~18–22 parole al secondo: leggibile. Una 4090 (1008 GB/s) fa ~50: fluidissimo. Due schede identiche in potenza ma con banda diversa possono avere prestazioni *diverse a metà*.

I tre profili, con i prezzi di novembre 2026

Profilo	Hardware	Prezzo indicativo	Cosa fa bene
Punto di partenza	RTX 4060 Ti 16 GB (o RTX 3060 12 GB, 455 €)	455–1.050 €	Modelli fino a 8B in Q4, 1–2 utenti, compiti di riassiunti/chat
Ufficio	RTX 4070 Ti Super 16 GB	~1.365 €	8–14B con contesto lungo, più utenti, velocità fluida
Team	RTX 4090 24 GB / RTX 5090 32 GB	5.805–7.235 €	30B+ e agenti in parallelo, contesti molto lunghi, multi-utente

E il Mac? Un Mac con Apple Silicon da 36 GB (M4 Max) ospita modelli fino a 32B con grazia — la banda della memoria unificata è alta e il consumo è modesto. Per molti team è la via più semplice, senza assemblare un PC. I prezzi cambiano di mese in mese: il pianificatore di setup aggiorna i numeri ogni giorno.

> **CPU sola:** se il budget è molto contenuto, una CPU moderna con 64 GB di RAM (banda ~40–90 GB/s) fa girare un 8B a 5–8 parole al secondo: lento, ma funziona per i riassiunti non in tempo reale. È il piano B, non il piano A.

4

Modelli e quantizzazione: come scegliere

Famiglie da conoscere

- **Qwen** — la famiglia open con il miglior supporto multilingua oggi, italiano compreso. È il nostro punto di partenza consigliato per contenuti in italiano.
- **Llama** (Meta) — la famiglia generalista più diffusa; 8B e 70B sono i punti di riferimento per dimensione.
- **Mistral** — efficienti e compatti, ottimi quando la macchina è piccola.
- **Qwen3-30B-A3B** — modello "misto" (30B totali, 3B attivi): qualità da modello grande, velocità da modello piccolo. La scelta più interessante per il profilo "team".

Quantizzazione: il compromesso qualità/peso

La quantizzazione comprime i pesi del modello da 16 bit (F16) a pochi bit, riducendo VRAM e velocità con un calo di qualità controllato. In pratica:

Quantizzazione	Bit/peso	Qualità	Quando usarla
Q4_K_M	4,8	≈ 85%	Il default: il miglior compromesso per quasi tutti i casi
Q5_K_M	5,7	≈ 92%	Se la VRAM lo permette e la qualità conta
Q6_K	6,6	≈ 95%	Modelli piccoli su hardware grande
Q2_K	3,5	≈ 55%	Solo se non entra altrimenti: la qualità crolla

Il metodo di scelta in tre mosse

- 1 **Parte dalla VRAM disponibile.** Sottrai ~2–4 GB di memoria di lavoro (dipende dal contesto) e dividiti: quello è il peso massimo del modello che entra.
- 2 **Scendi di famiglia, non di qualità.** Meglio un 8B in Q5 che un 14B in Q2: la qualità della quantizzazione pesa quanto la dimensione.
- 3 **Prova con il tuo compito reale.** Il benchmark generale non basta: metti 10 prompt del tuo lavoro e confronta. È l'unico test che conta.

5

Strumenti: Ollama, Open WebUI, vLLM

Il modello è solo i pesi. A farlo girare serve un motore di inferenza, e a fargli parlare serve un'interfaccia. Tre strumenti coprono il 95% dei casi.

Ollama — il motore per iniziare

Un'unica riga di installazione, modelli che si scaricano con un comando, API locale pronta. È il punto di partenza: se devi "far girare qualcosa oggi", è Ollama. Gestisce un modello alla volta a utente, il che è esattamente ciò che serve a 1–5 persone.

Open WebUI — l'interfaccia

Un'interfaccia web in stile chat sopra Ollama (o altre API): conversazioni salvate, più utenti, upload di documenti per il RAG (fare domande ai tuoi documenti). In 10 minuti di setup hai il "ChatGPT di casa", sul tuo server.

vLLM — quando diventa produzione

Quando i modelli servono a molti utenti in parallelo con contesti lunghi, vLLM fa il batching delle richieste e usa la VRAM in modo più efficiente. È lo strumento "server": più complesso da impostare, ma è la risposta quando Ollama inizia a fare la coda.

> **La progressione naturale:** Ollama (oggi) → Open WebUI (questa settimana) → vLLM (quando il team cresce). Non partire da vLLM: la maggior parte delle attività non lo incontra mai, e se lo incontra, lo fa con esperienza alle spalle.

Strumento	Setup	Utenti	Quando
Ollama	~5 min	1–5	Sempre il punto di partenza
Open WebUI	~10 min	2–15	Quando serve l'interfaccia di squadra
vLLM	ore	10+	Quando le richieste si accumulano in coda

6

Primi passi: 10 minuti, due comandi

Su un Mac o un Linux (e su Windows, via WSL2), il percorso da zero a prima risposta:

```
# 1. Installa Ollama
curl -fsSL https://ollama.com/install.sh | sh

# 2. Scarica un modello 8B (Q4_K_M, ~4,9 GB)
ollama pull qwen2.5:7b

# 3. Parla
ollama run qwen2.5:7b "Riassumi in 3 frasi: <il tuo testo>"
```

Sul Mac l'installazione è un .dmg; il modello si sceglie con la stessa sintassi e gira sulla memoria unificata. Su un PC con NVIDIA, Ollama rileva la GPU da solo.

Poi, in quest'ordine

- 1 **Definisci il caso d'uso.** Prendine uno solo. Scrivi i 10 prompt reali che gli daresti.
- 2 **Misura.** Velocità (parole/secondo), qualità (su quei 10 prompt), memoria occupata. Annota i numeri.
- 3 **Confronta con il pianificatore.** Sul pianificatore di setup inserisci attività, utenti e velocità: ti dice l'hardware minimo e i prezzi attuali, e li confronta con i tuoi numeri.
- 4 **Aggiungi Open WebUI** quando il secondo collega vuole provarci. Da lì in poi è squadra.

> **Attenzione al contesto:** se devi incollare documenti interi, verifica quanti token entrano (l'8B di Qwen2.5 arriva a 128K di contesto teorico, ma la VRAM decide quello *pratico*). Per i documenti: il RAG, che è l'argomento del numero più recente della newsletter di iaLocale.

7

I limiti: quando l'IA locale non basta

Siamo onesti, perché è l'unico modo per scegliere bene.

- **Frontier gap.** Per i task di ragionamento più profondi, i modelli cloud più grandi restano davanti. Se il tuo caso è quello, la risposta è ibrida: locale per il 90% dei volumi, cloud per il 10% difficile.
- **Il primo costo è reale.** 300–800 euro di hardware non sono niente, ma non sono zero. Sotto un certo volume di utilizzo, il cloud resta più economico.
- **I documenti lunghi hanno un prezzo.** Ogni parola di contesto occupa VRAM. I contesti molto lunghi (centinaia di pagine) richiedono hardware serio o tecniche di RAG.
- **I modelli cambiano.** Oggi il miglior 8B di italiano tra 6 mesi potrebbe essere un altro. La buona notizia: con l'infrastruttura locale, cambiare modello è un comando, non un progetto.

> **La regola pratica:** se i tuoi dati escono dal cloud oggi senza problemi e i volumi sono bassi, parti con un servizio gestito e tieni questa guida come roadmap. Nessun fornitore in particolare: la scelta dipende da budget, privacy e tipo di modello — il catalogo dei servizi di inferenza raccoglie i principali, con i dettagli di ciascuno. Se i dati sono sensibili, i volumi crescono o la disponibilità conta, l'ordine si inverte: la via locale non è più un optional.

E quando serve di più?

La scala esiste: da una 4090 a un server multi-GPU, passando dai Mac da 128 GB. Il pianificatore di setup mostra la scala completa con i prezzi, così la crescita è un piano, non una scommessa.

8

Dove andare avanti

Questa guida è il "cosa" e il "perché". Il "come" dettaglio per dettaglio è già sul sito:

- **Il pianificatore di setup** — inserisci attività, utenti, velocità: ottieni l'hardware minimo, i prezzi aggiornati ogni giorno e i link alle fonti. /it/strumenti/setup
- **5 casi d'uso pratici per la tua attività** — i casi concreti con numeri e flussi. /it/blog/ia-locale-attivita-5-casi-duso-pratici
- **RAG: fai domande ai tuoi documenti** — la tecnica per i contesti lunghi. /it/blog/rag-fai-domande-ai-tuoi-documenti
- **GPU o CPU: quando la CPU basta** — la regola dei 40 GB/s. /it/blog/gpu-o-cpu-quando-la-cpu-basta
- **La newsletter** — ogni numero è un passo, con l'articolo completo. /it/newsletter (archivio: /it/newsletter/archivio)
- **La pagina aziende** — questa stessa ragionamento, con tabelle e costi aggiornati. /it/aziende

> **Il metodo, in una frase:** parti da un caso d'uso reale, misura con i tuoi prompt, confronta con i numeri del pianificatore e scala solo quando i numeri lo chiedono. È tutta la filosofia di iaLocale: l'IA locale non è una scelta ideologica, è un calcolo — e i calcoli si fanno.

Guida all'IA Locale per Aziende · 1ª edizione, settembre 2026 · iaLocale Team
Prezzi di riferimento: mercato europeo, novembre 2026 — variano nel tempo; verifica sempre il pianificatore di setup per i valori correnti. Ultimo aggiornamento: 22 novembre 2026.
Documento generato con IA.