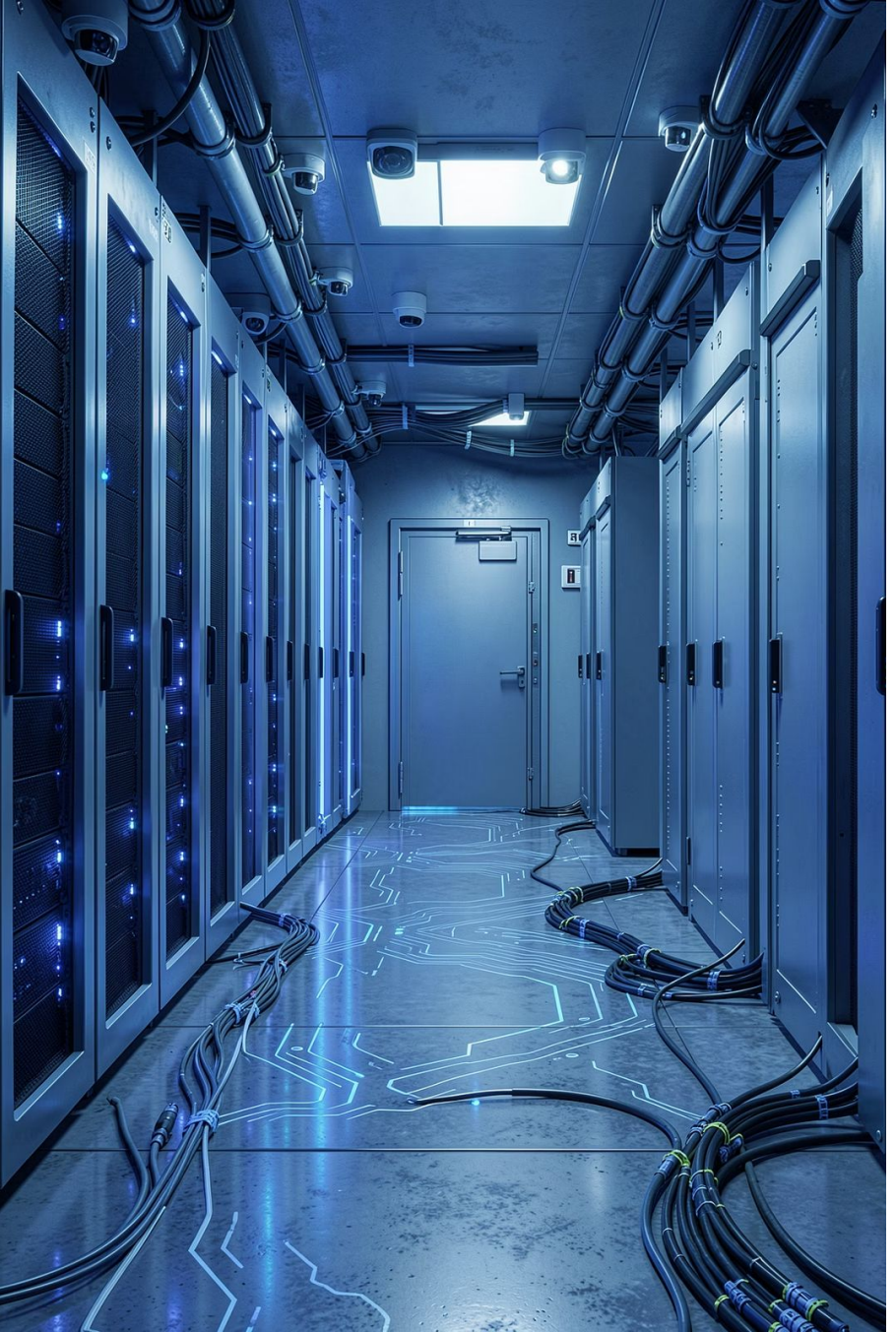
Gruppo BitAgorà Smart Network

Quanto hardware serve per un LLM in locale

Senza tirare a indovinare: una guida pratica per valutare se il proprio hardware regge un modello linguistico prima ancora di scaricare un singolo file.



Perché eseguire un LLM in locale?

🔒 Privacy

I dati aziendali restano all'interno della propria infrastruttura.

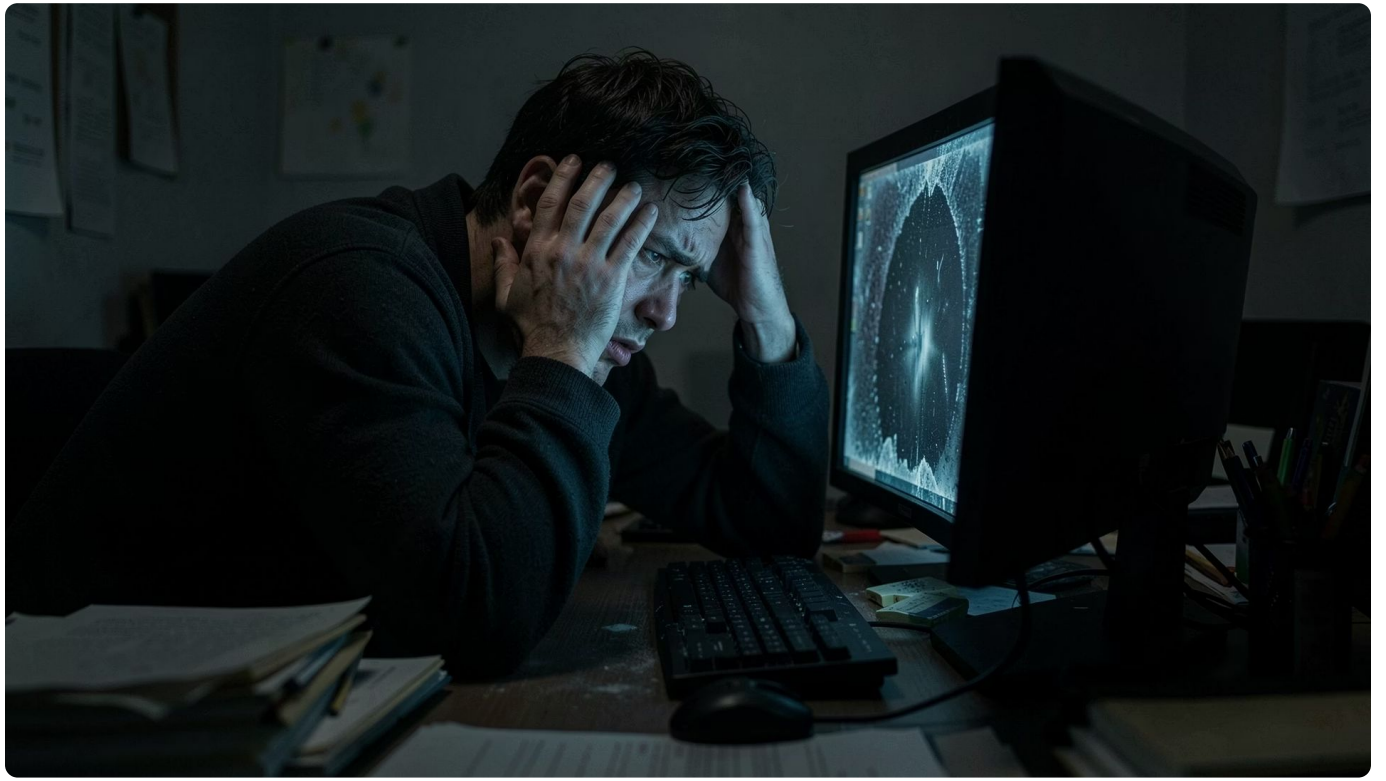
💰 Costi

Nessun costo per ogni singola query inviata al modello.

🔗 Indipendenza

Nessuna dipendenza da servizi cloud che possono cambiare prezzi o condizioni.

Il problema: come saperlo prima?



Per un po' la risposta è arrivata **per tentativi**: installare, provare, aspettare che il sistema rallentasse fino a diventare inutilizzabile o andasse in errore per mancanza di memoria video.

Oggi esiste uno strumento gratuito che fa questi conti **prima ancora di scaricare un file**: l'**ApX VRAM Calculator** su *apxml.com/tools/vram-calculator*.

Quattro concetti chiave

VRAM

Memoria della scheda video. Il modello deve starci quasi interamente per funzionare in modo fluido.

Parametri

La "taglia" del modello (es. 7B, 14B, 70B). Più parametri = più capacità, ma anche più memoria richiesta.

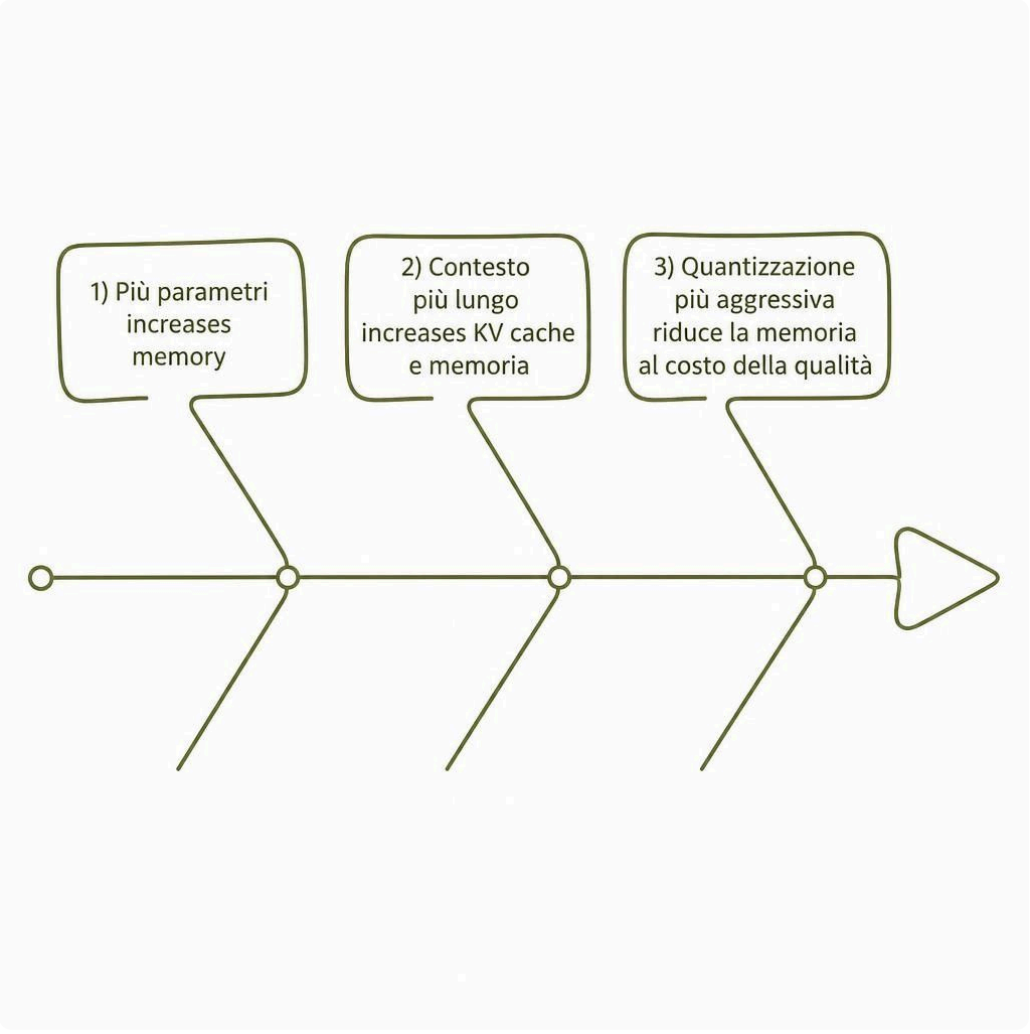
Contesto

Quanto testo il modello tiene "in mente", misurato in token. Più contesto = più memoria occupata.

Quantizzazione

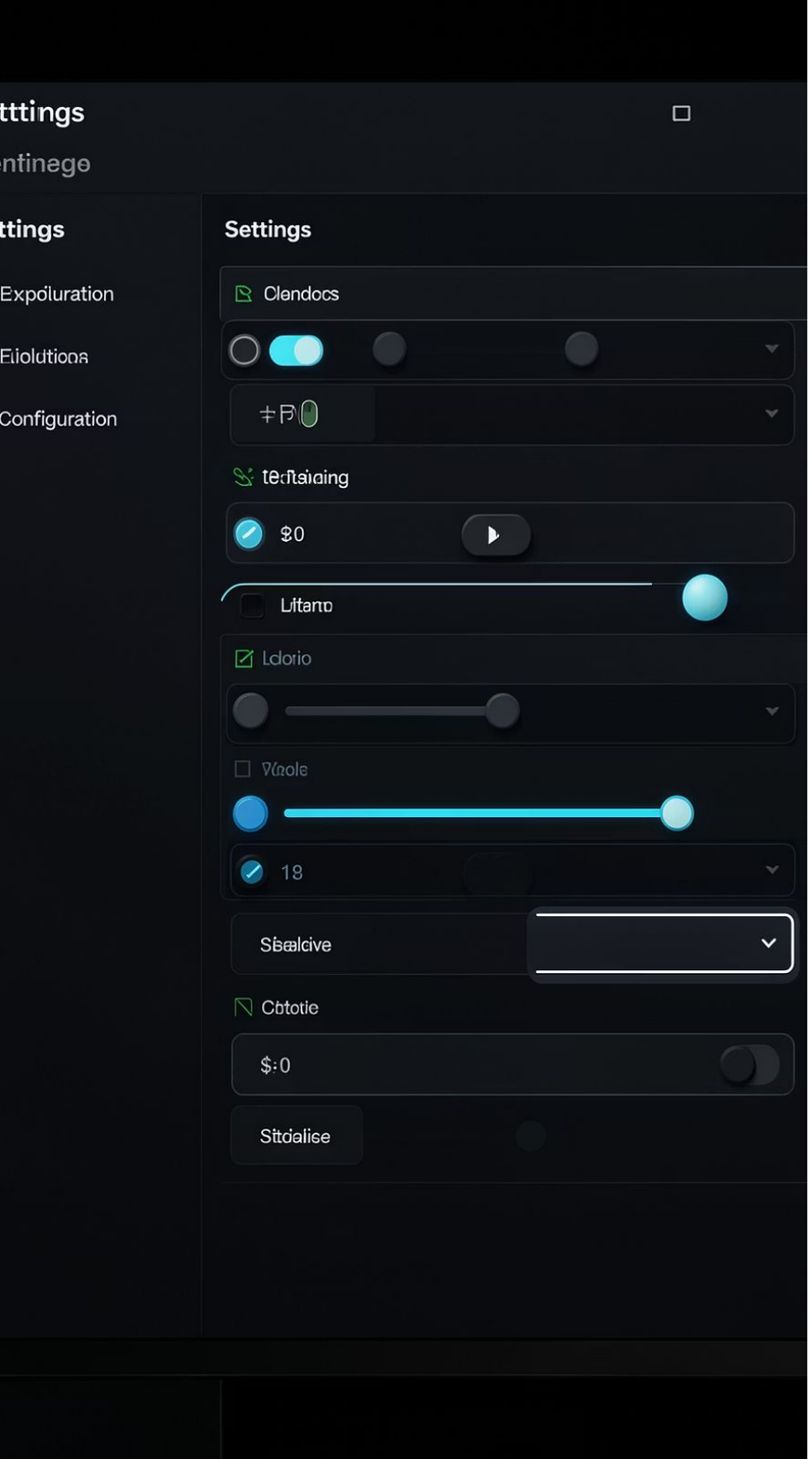
Compressione dei numeri interni del modello: meno memoria, ma con un compromesso sulla precisione.

La logica dietro i numeri



Il calcolatore stima la VRAM necessaria incrociando tre fattori che si muovono in direzioni diverse:

- **Parametri maggiori** → più memoria richiesta
- **Contesto più lungo** → KV cache più grande → più memoria
- **Quantizzazione aggressiva** → meno memoria, ma qualità ridotta



La configurazione di partenza consigliata

01

Modalità Inference

Non Fine-tuning, che serve solo per addestrare modelli da zero.

02

Batch size 1, utente singolo

Una domanda alla volta: il caso d'uso più comune in locale.

03

Contesto a 8K token

Un valore medio, adatto alla maggior parte degli scenari aziendali.

04

GPU reale installata

Usare il modello esatto della propria scheda: anche poca VRAM in meno può ribaltare l'esito.

Come leggere il risultato

✅ Ampio margine

La VRAM richiesta resta ben sotto quella disponibile. Si può procedere.

⚠️ Al limite

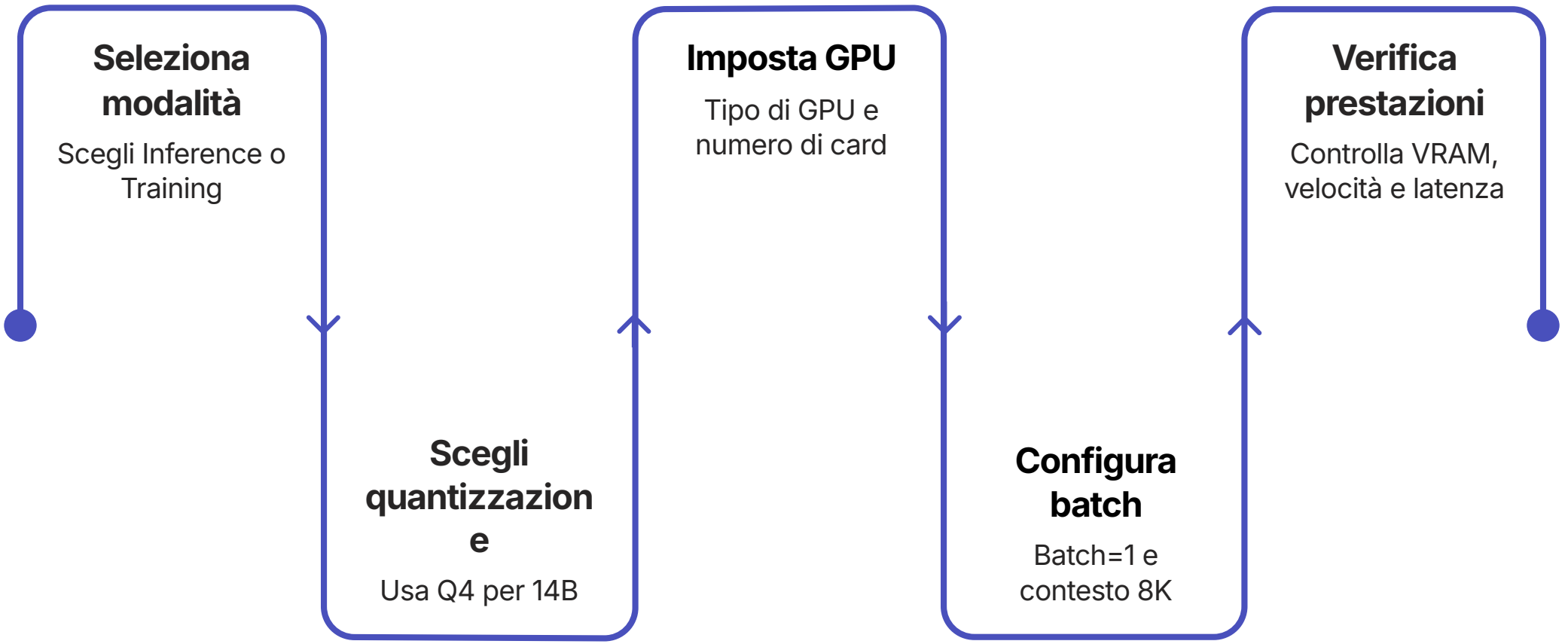
Il modello entra, ma un contesto più lungo o altri programmi aperti possono saturare la GPU.

❌ Non entra

Serve quantizzazione più spinta, contesto ridotto, più VRAM, più schede o offloading.

ⓘ Attenzione: un modello che entra in memoria non è ancora un modello piacevole da usare. Contano anche i **token al secondo** generati e la **latenza** prima della prima risposta.

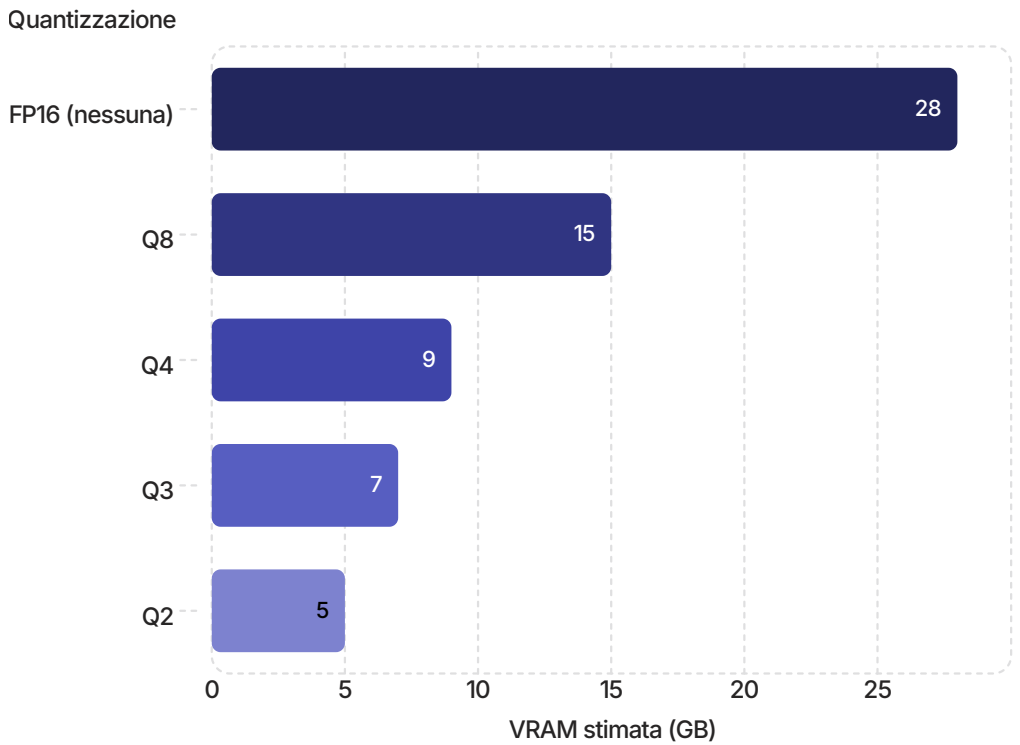
Un esempio concreto: modello da 14B



Resta comunque una stima: runtime, driver, architettura del modello e altri carichi sulla macchina possono spostare il risultato reale.

VRAM: quanto pesa davvero la quantizzazione?

La scelta del livello di quantizzazione ha un impatto diretto sulla memoria richiesta. Ecco come varia per un modello da 14B parametri (valori indicativi).



Il punto di partenza per ogni decisione hardware

Se in azienda ci si sta chiedendo se ha senso portare un LLM in locale invece di affidarsi al cloud, **questi numeri sono il primo posto da cui partire** prima di acquistare qualsiasi hardware.

🧱 Calcola prima

Usa l'ApX VRAM Calculator su apxml.com

⚙️ Configura bene

Inference, GPU reale, 8K contesto, Q4 come punto di partenza

📊 Valuta tutto

VRAM, token/sec e latenza: tutti e tre contano per l'usabilità reale