STUDIO · SPARKTEST

Modelli e hardware per SparkTest on premise

INDICE

Sintesi 3
Che cosa abbiamo trovato 3
Tre classi di configurazione 5

1. Che cosa SparkTest chiede a un modello 6
1.1 Dove viene chiamato il modello 6
1.2 Dimensioni di prompt e risposte 7
1.3 Chiamate per test e per notte 7
1.4 Latenza e concorrenza 8
1.5 Requisiti minimi del modello 9

2. Criteri di selezione 9

3. I modelli 10
3.1 Candidati per un singolo nodo 10
3.2 Modelli da data center 11
3.3 Condizioni di licenza che richiedono una decisione 13
3.4 Note sul panorama 13

4. L'hardware 14
4.1 Classe 1: sistemi compatti a memoria unificata 14
4.2 Classe 2: GPU da workstation 15
4.3 Classe 3: acceleratori e server da data center 15
4.4 Che cosa implicano le tabelle dell'hardware 17

5. Ingombro e velocità 18
5.1 Memoria per modello (stima) 18
5.2 Matrice di ingombro (stima) 19
5.3 Velocità: stime accanto alle misure pubbliche 21
5.4 Concorrenza 22
5.5 Throughput da data center (misurato, nessuna stima) 23
5.6 Misure dell'agente da un capo all'altro 23
5.7 Misurato sui compiti di SparkTest 24

6. Configurazioni confrontate e consigliate 25
6.1 Classe 1: sistema compatto a memoria unificata 28
6.2 Classe 2: singola GPU da workstation 28
6.3 Classe 3: server multi-GPU 29
6.4 Scelta per taglia 29

7. Portare il proprio modello 30
7.1 Requisiti 30
7.2 Collegare l'endpoint 30
7.3 Software di serving: compatibilità al 16 e al 17 novembre 2026 32
7.4 Lista di controllo per il reparto IT del cliente 33
7.5 Qualifica con un insieme di test 33

8. Rischi e limiti 34
8.1 Benchmark dichiarati e chi li ha misurati 34
8.2 Saturazione e contaminazione 34
8.3 Perdita di qualità con la quantizzazione 34
8.4 Licenze 34
8.5 Prezzi e disponibilità 34
8.6 Limiti delle stime e delle misure 35
8.7 Italiano 35

9. Metodo e riproducibilità 35
9.1 Fonti ed etichette 35

9.2 Formule 36
9.3 Valori di taratura (calcolati dal calcolatore da `dati/measurements.json`) 36
9.4 Ipotesi 37
9.5 File e come rieseguire 37

Riferimenti 37
Carico di SparkTest e software di serving (nota `fonti/workload-serving.md`) 37
Modelli (nota `fonti/models.md`) 40
Hardware (nota `fonti/hardware.md`) 43
Misure di inferenza (nota `fonti/measurements.md`) 46
Benchmark (nota `fonti/benchmarks.md`) 48
Riferimenti di metodo aggiunti dall'autore 49

Verifica 49

Studio sui modelli linguistici open-weight, sull'hardware di inferenza e sugli endpoint di modello gestiti dal cliente

SparkTech, Torino. Data dello studio: 16 novembre 2026; rivisto il 17 novembre 2026 dopo una verifica indipendente dei suoi dati, poi aggiornato lo stesso giorno per le modifiche al prodotto del 17 novembre 2026 e con la prima misura di SparkTech dei compiti di SparkTest su modelli open-weight (sezione 5.7; vedi "Verifica" in fondo). Tutti i dati sono aggiornati al 16 novembre 2026, salvo dove è indicata un'altra data; ogni fonte web è stata consultata in quella data, salvo diversa indicazione nel riferimento.

Come leggere lo studio. I codici tra parentesi quadre, come [models-12] o [hardware-9], rimandano ai riferimenti numerati in fondo. Ogni dato ha una fonte, e le tabelle e la sintesi indicano di che tipo di dato si tratta:

- **dichiarato**: affermato dal produttore (sviluppatore del modello, venditore di hardware, progetto software);
- **misurato**: misurato da qualcuno diverso da questo studio; il testo dice chi: una parte indipendente (analista, consorzio di benchmark, stampa, comunità), lo sviluppatore di un runtime oppure chi sottopone un risultato MLPerf (risultati sottoposti a revisione tra pari da MLCommons);
- **osservato**: un prezzo letto su una pagina pubblica datata;
- **stima**: calcolato dal calcolatore di questo studio (`calcolatore.mjs`), mai una misura;
- **misurato da SparkTech**: misurato da SparkTech sui compiti di SparkTest, su un server di prova, per il documento di accompagnamento riassunto nella sezione 5.7; mai la velocità di un nodo.

I dati su SparkTest vengono dal suo codice sorgente al 16 novembre 2026, aggiornati per le modifiche del 17 novembre 2026 (riferimenti interni da [workload-serving-1] a [workload-serving-10] e [workload-serving-67], senza URL pubblico); le chiamate e i token per notte ricavati da questi dati sono stime. Dove un dato è stato cercato e non trovato, il testo dice "non trovato".

Termini. Un *token* è un frammento di testo, circa quattro caratteri in inglese. Il *prefill* è la lettura del prompt e il *decode* la scrittura della risposta, entrambi in token al secondo (tok/s). La *cache KV* è la memoria che serve a ogni richiesta aperta; una *sessione da 32k* è una richiesta fino a 32.768 token (le lunghezze di contesto usano k = 1.024 token). Un modello *mixture-of-experts (MoE)* usa per ogni token solo una parte dei suoi parametri, i *parametri attivi*; un modello *denso* li usa tutti. *BF16, 8 bit, 4 bit* indicano quanti bit servono per memorizzare ogni peso: meno bit richiedono meno memoria e sono più veloci, con una possibile perdita di qualità (sezione 8.3). Una *raffica* è l'arrivo contemporaneo di tutte le richieste concorrenti di una taglia. L'*obiettivo* è il tempo a cui punta questo studio; il *limite rigido* è il tempo oltre il quale l'agente interrompe una chiamata.

Sintesi

Che cosa abbiamo trovato

01 **SparkTest chiede poco a un modello in volume e molto in latenza.** Un'esecuzione il cui piano di test regge ancora non chiama il modello. Il modello lavora quando un test viene scritto o modificato (da 8 a 16 chiamate con prompt di circa 4.300-4.900 token e risposte di 60-120 token), quando un'esecuzione fallisce (1 o 2 chiamate di riparazione con prompt di circa 3.200-3.700 token e risposte di 440-720 token, poi una diagnosi se il fallimento persiste) e su richiesta [workload-serving-1][workload-serving-4][workload-serving-10]. Un test fallito non viene codificato di nuovo: dal 17 novembre 2026 la sua esecuzione successiva esegue lo stesso piano e lo ripara sotto il controllo dello scenario dell'agente [workload-serving-67]. Una notte tipica con 400 test mantenuti richiede circa 54-106 chiamate; una notte di rilascio con 1.500 test circa 590-1.160 chiamate e 2,2-5,1 milioni di token di prompt (stima, sezione 1.3) [workload-serving-10]. Il vincolo è il tempo di ogni chiamata mentre sono in corso 2, 4 o 8 esecuzioni parallele (taglie S, M, L) e un'analisi API: l'agente interrompe una chiamata di codifica o di riparazione a 60 s e una diagnosi a 30 s [workload-serving-1][workload-serving-4][workload-serving-9].

02 **Con queste latenze, su un sistema compatto entra una sola classe di modelli.** I modelli MoE con circa 3B di parametri attivi e attenzione ibrida (Qwen3.6-35B-A3B, NVIDIA Nemotron 3.5 Lightning 30B-A3B, poolside Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b) richiedono da 12 a 20 GB di pesi a 4 bit e da 0,2 a 1,4 GB di cache KV per sessione da 32k (stima, sezione 5.1). I modelli densi da 24B a 32B generano da 6 a 12 tok/s sui sistemi compatti a memoria unificata (misurato dallo sviluppatore di un runtime e dalla comunità [measurements-11][measurements-20][measurements-21]) e vanno su una GPU da workstation. I modelli da 117B a 125B di parametri in memoria (gpt-oss-120b, Mistral Small 4, Nemotron 3.5 Super VL, Qwen3.5-122B-A10B) entrano in 128 GB a 4 bit, ma nessuno rispetta gli obiettivi su un sistema compatto: gpt-oss-120b e Mistral Small 4 restano entro i limiti rigidi dell'agente fino alla taglia M

su GB10, e su Ryzen AI Max+ 395 solo gpt-oss-120b, fino alla taglia S; una GPU da workstation da 96 GB li porta entro gli obiettivi (stima, sezione 6). I modelli da data center (da 180B a 2,8T di parametri) richiedono un server multi-GPU.

03 **Le licenze sono il primo filtro.** Apache-2.0, MIT e OpenMDW-1.1 coprono la maggior parte dei candidati per un singolo nodo. Llama 4 non concede diritti sui suoi modelli multimodali alle società con sede nell'UE; ERNIE 5.5 Flash Base esclude l'UE; MiniMax M2.7 è non commerciale; Mistral Medium 3.5 e Devstral 2 non concedono diritti oltre USD 20 milioni di ricavi mensili della società; Qwen3.8-Flash-Next richiede una licenza separata per i prodotti "AI Work Assistant" [models-44][models-45][models-162] [models-146][models-59][models-61][models-22]. Chi sia il licenziatario quando il modello gira su un nodo nella rete del cliente è una questione legale a cui questo studio non risponde.

04 **Nessuna evidenza pubblica misura il compito di SparkTest.** Sull'Artificial Analysis Intelligence Index v5.0.1 il primo modello open-weight in classifica ottiene 46 contro 65 del primo modello proprietario; i modelli per un singolo nodo hanno pochi punteggi (Qwen3.8-27B: 20; gpt-oss-120b e Gemma 4 31B: 4) (indipendente) [benchmarks-1][benchmarks-3][benchmarks-4][benchmarks-5]. Questo studio non ha trovato una classifica indipendente sull'aderenza agli schemi JSON né una valutazione specifica in italiano dei modelli open-weight del 2026. Per lo stesso modello, i dati del produttore e quelli indipendenti differiscono fino a 36 punti [benchmarks-49][benchmarks-18][benchmarks-23]. Un modello va qualificato sull'insieme di test di SparkTest prima dell'uso (sezione 7.5); la prima misura di questo tipo fatta da SparkTech è il punto 7.

05 **La banda limita la generazione; il calcolo limita la lettura del prompt.** Le due piattaforme compatte confrontate hanno 256 GB/s (ricavati dal bus di memoria dichiarato) e 273 GB/s (dichiarati) e generano a velocità simili, ma leggono i prompt a velocità diverse: il prefill di gpt-oss-120b a 32k di profondità è di 1.567 tok/s su DGX Spark (misurato dallo sviluppatore del runtime) contro 234-653 tok/s su Ryzen AI Max+ 395 secondo il backend software (comunità) [measurements-2][measurements-20]; in MLPerf Inference v6.1 Edge Agentic un turno dell'agente ha richiesto 3,81 s contro 7,06 s (misurato da chi ha sottoposto il risultato, revisione tra pari) [measurements-26]. I prompt di SparkTest sono lunghi e le risposte brevi, quindi la lettura del prompt conta.

06 **I prezzi sono volatili.** Prezzo di listino di DGX Spark: USD 3.999 (ottobre 2025), 4.699 (febbraio 2026), 6.950 (ottobre 2026). Una GeForce RTX 5090 era in vendita in Germania da EUR 7.399,99 il 16 novembre 2026, contro un listino UE di EUR 2.229 (osservato) [hardware-3][hardware-4][hardware-5][hardware-44][hardware-45]. Ogni prezzo qui riportato è un'indicazione datata.

07 **Misurato sui compiti di SparkTest (misurato da SparkTech).** Su un server di prova senza GPU, Qwen3.6-35B-A3B (Q4_K_M) e gpt-oss-20b (MXFP4), serviti da llama.cpp, hanno risposto agli otto compiti del kit di qualificazione di SparkTest in inglese e in italiano, con i prompt e i controlli dell'agente. gpt-oss-20b ha superato tutti i 23 controlli obbligatori, Qwen3.6 22 (un piano respinto dal validatore dell'agente); ogni risposta vincolata da uno schema ha superato il controllo dello schema; entrambe le diagnosi in italiano hanno citato le prove. Entrambi i modelli hanno cambiato un valore atteso quando un prompt di riparazione li spingeva a farlo, e il controllo dello scenario dell'agente ha respinto le due riparazioni, quindi la regressione è rimasta visibile. I tempi (mediana di 101 e 108 s per chiamata, 1 chiamata su 8 entro i limiti dell'agente) descrivono un processore da server del 2016, non un nodo: la lettura del prompt, che è ciò che manca a una CPU, ha occupato la maggior parte di ogni chiamata (sezione 5.7).

Tre classi di configurazione

Taglie: S = 100 test mantenuti e 2 esecuzioni parallele, M = 400 e 4, L = 1.500 e 8 [workload-serving-9]. I verdetti che seguono sono stime (sezioni 5 e 6) per una raffica in cui arrivano insieme tutte le chiamate concorrenti della taglia, con le sessioni da 32k concorrenti in memoria:

- **entro gli obiettivi**: la raffica rispetta gli obiettivi di questo studio (giro di codifica 15 s, riparazione 30 s, diagnosi 30 s);
- **entro i limiti rigidi**: la raffica rispetta i limiti rigidi dell'agente (60 s, 60 s, 30 s); SparkTest funziona, più lentamente (sezione 6);
- **utilizzabile fino a**: la taglia più grande entro i limiti rigidi sia con i prompt tipici (giro di codifica da 4.941 token) sia con la documentazione dell'applicazione al limite di 24.000 caratteri (giro di codifica da 11.093 token).

CLASSE	HARDWARE CONFRONTATO	MEMORIA; BANDA	MODELLI ADATTI ALLA CLASSE (STIMA)	TAGLIE (STIMA)	PREZZO INDICATIVO, 16 NOV 2026 (OSSERVATO)	CHE COSA NON FA
1. Sistema compatto a memoria unificata	sistemi NVIDIA GB10 (DGX Spark e sistemi dei partner); sistemi AMD Ryzen AI Max+ 395 e PRO 495	da 64 a 192 GB; 256 GB/s (ricavati) o 273 GB/s (dichiarati)	MoE con circa 3B attivi a 4 bit: Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b	GB10 con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1: entro gli obiettivi fino a M con i prompt tipici, S con la documentazione al limite; utilizzabile fino a L. GB10 con Gemma 4 26B-A4B o gpt-oss-20b: obiettivi fino a S, utilizzabile fino a L. Ryzen AI Max+ 395: nessuna taglia entro gli obiettivi; utilizzabile fino a M (fino a S con Gemma 4 26B-A4B o gpt-oss-20b)	da EUR 3.899,96 (Ryzen AI Max+ 395, 128 GB) e EUR 5.999 (GB10, 128 GB), prezzo di strada DE [hardware-22] [hardware-9]	la taglia L entro gli obiettivi; i modelli densi da 24B a 32B entro gli obiettivi; i MoE da 117B a 125B entro gli obiettivi (entro i limiti rigidi: fino a M su GB10; su Ryzen AI Max+ 395 solo gpt-oss-120b, fino a S); i modelli da data center
2. Singola GPU da workstation	RTX PRO 6000 (96 GB), RTX PRO 5000 (48 o 72 GB), schede da 32 GB (RTX PRO 4500, Radeon AI PRO R9700, Arc Pro B70; GeForce RTX 5090 con la restrizione di licenza della sezione 4.2)	da 32 a 96 GB per scheda; da 608 a 1.792 GB/s (dichiarati)	96 GB: tutti i candidati per un singolo nodo; da 48 a 72 GB: MoE con 3B attivi da 4 a 8 bit, densi da 27B a 31B da 4 a 8 bit; 32 GB: MoE con 3B attivi a 4 bit	96 GB: entro gli obiettivi fino a L per i MoE con 3B attivi, gpt-oss-120b e Mistral Small 4 (gli ultimi due: M con la documentazione al limite); denso da 27B a 8 bit: M. Da 48 a 72 GB: L per i MoE con 3B attivi (M con la documentazione al limite). 32 GB: M per Qwen3.6-35B-A3B, Nemotron 3.5 Lightning e gpt-oss-20b (S con la documentazione al limite), utilizzabile fino a L; RTX 5090: L	scheda da EUR 1.659 (Arc Pro B70) a EUR 16.489,98 (RTX PRO 6000 Max-Q), prezzo di strada DE; computer ospite escluso [hardware-61][hardware-48]	i MoE da 117B a 125B sotto i 96 GB; i modelli da data center; diverse schede hanno restrizioni per l'uso in data center (sezione 4.2)
3. Server multi-GPU	4 x RTX PRO 6000 Server Edition; 8 x H100, H200 o B200 (HGX); 8 x Instinct MI355X	da 384 GB a 2,3 TB	modelli da data center: Qwen3.5-397B-A17B, GLM-5.3-Flash, DeepSeek-V4-Flash, Mistral Large 3, Nemotron 3 Ultra; GLM-5.3 e Kimi K3.1 Pro con licenze proprie	non limitato dal carico di SparkTest (throughput misurato, sezione 5.5)	da USD 195.000 (nodo 8 x H100 usato) a USD 470.000 (nodo 8 x B200 nuovo), USA, ottobre 2026 [hardware-87] [hardware-88]	non è un'appliance; costo e consumo (fino a circa 14 kW per un sistema DGX B200 o DGX B300, dichiarati) sono di un altro ordine [hardware-68][hardware-69]

Portare il proprio modello. Un cliente che gestisce già un modello può collegarlo tramite un endpoint compatibile OpenAI (`LLM_URL`). Servono: imposizione dello schema JSON lato server, almeno 32k token di contesto, ragionamento disattivato, tante richieste concorrenti quante sono le esecuzioni parallele della taglia più una, e tempi per chiamata entro gli obiettivi. Fino al 16 novembre 2026 la richiesta dell'agente era accettata da llama.cpp server e da Ollama, mentre vLLM 0.33 e SGLang 0.5.23 la rifiutavano perché non portava il nome dello schema. Dal 17 novembre 2026 l'agente dà un nome al suo schema e chiede di non ragionare anche con `chat_template_kwargs` e `reasoning_effort` , i campi che questi server leggono; stando al loro codice sorgente vLLM e SGLang ora accettano la richiesta, ma nessuna esecuzione contro di loro è stata ancora fatta [workload-serving-16][workload-serving-24][workload-serving-31][workload-serving-40][workload-serving-67]. La sezione 7 elenca i requisiti, le modifiche ancora aperte e una procedura di qualifica.

Che cosa significa per il vostro reparto IT

- Se il nodo lo fornisce SparkTech, SparkTech dimensiona ogni SparkNode sui vostri volumi con il metodo della sezione 9 e lo qualifica sui vostri test prima dell'uso (sezione 7.5).
- Se portate il vostro modello, verificate l'endpoint con la sezione 7.1 e con la lista di controllo della sezione 7.4, poi eseguite su di esso il kit di qualificazione di SparkTest (sezioni 5.7 e 7.5). Dal 17 novembre 2026 la richiesta dell'agente porta il nome dello schema che vLLM, SGLang e NVIDIA NIM richiedono; OpenVINO Model Server e TGI richiedono ancora le modifiche elencate nella sezione 7.3.
- Prevedete 3, 5 o 9 richieste contemporanee (taglie S, M, L), ognuna con un prompt da 4.000 a 11.000 token. Il volume notturno è piccolo; conta il tempo di ogni chiamata.
- Entro gli obiettivi (stima): un sistema GB10 fino alla taglia M con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1, una scheda da workstation da 32 GB fino a M, una RTX PRO 5000 o RTX PRO 6000 fino a L. Dove la documentazione dell'applicazione arriva al limite, GB10 e le schede da 32 GB scendono a S e la RTX PRO 5000 a M. Confermate in sede con la sezione 7.5.
- Stabilite prima dell'acquisto chi è titolare della licenza del modello presso di voi (sezione 8.4).

1. Che cosa SparkTest chiede a un modello

SparkTest trasforma un test scritto in linguaggio naturale in un piano di test JSON, esegue il piano in un browser e lo ripara quando l'applicazione cambia. Questa sezione descrive il carico sul modello così come risulta dal codice; non dipende da alcun modello. Dati di dettaglio e metodo: nota di accompagnamento `fonti/workload-serving.md` e dati `dati/workload.json` .

1.1 Dove viene chiamato il modello

OPERAZIONE	QUANDO AVVIENE	RISPOSTA	SCHEMA IMPOSTO DAL SERVER	TIMEOUT PER CHIAMATA	FONTI
Giro di codifica	test nuovo o modificato, oppure uno script salvato senza piano; dal 17 novembre 2026 mai dopo un'esecuzione terminata "failed" (il piano viene eseguito di nuovo e riparato)	JSON: fino a 5 passi successivi e una nota	sì	60 s; esplorazione completa 4 min, al massimo 15 giri	[workload-serving-1] [workload-serving-4] [workload-serving-67]
Codifica in un colpo solo	browser non disponibile o budget di esplorazione esaurito	piano JSON completo (da 1 a 60 passi)	sì	60 s	[workload-serving-1] [workload-serving-2]
Riparazione	un'esecuzione è fallita per un motivo diverso dall'infrastruttura	piano JSON completo corretto	sì	60 s; fino a 2 tentativi per esecuzione	[workload-serving-1] [workload-serving-4]
Diagnosi	un'esecuzione termina "failed"	Markdown, da 60 a 120 parole, nella lingua del test	no	30 s	[workload-serving-1] [workload-serving-4]
Analisi API	su richiesta, per applicazione	JSON, al massimo 6 suggerimenti	sì	150 s	[workload-serving-1] [workload-serving-6]
Esecuzione semplice	suite notturna e su richiesta	nessuna	nessuno	nessuno	[workload-serving-4] [workload-serving-6]

Ogni piano viene validato dall'agente anche quando il server ha imposto lo schema. Una riparazione che toglie asserzioni o troppa parte dello scenario viene respinta, e dal 17 novembre 2026 lo è anche una riparazione che cambia un valore atteso (un testo o un indirizzo da trovare, un conteggio minimo, il tipo di una verifica di presenza) o un valore che il test digita (un campo, un filtro, la riga che apre): selettori, etichette e attese possono cambiare, ciò che il test verifica no. L'esecuzione allora fallisce come possibile regressione [workload-serving-1][workload-serving-2][workload-serving-67]. Nella versione esaminata, l'endpoint del cliente serve le operazioni dell'agente elencate sopra; i suggerimenti di test e la chat sui test della console usano una connessione al modello separata, che `LLM_URL` non configura, con prompt fino a circa 26.000 token [workload-serving-5][workload-serving-10]. Un'installazione on premise deve dichiarare da dove queste due funzioni prendono il modello (lo stesso endpoint locale, un altro modello locale, oppure disattivate), così che la verifica del perimetro dei dati della sezione 7.1 le comprenda.

1.2 Dimensioni di prompt e risposte

I caratteri sono stati misurati eseguendo i generatori di prompt dell'agente su input sintetici (nessun modello è stato chiamato). Token: caratteri divisi per 4,0 per l'inglese, la regola empirica di OpenAI [workload-serving-11], e per 3,5 per l'italiano, un'ipotesi dello studio sul carico [workload-serving-10].

OPERAZIONE	CASO	TOKEN DI PROMPT EN / IT	TOKEN DI RISPOSTA EN / IT	FONTI
Giro di codifica	test tipico da 20 passi, media di 11 giri	4.324 / 4.941	62 / 70 (JSON compatto)	[workload-serving-10]
Giro di codifica	documentazione dell'applicazione al limite di 24.000 caratteri	9.707 / 11.093	62 / 70	[workload-serving-10]
Riparazione	piano tipico da 20 passi	3.233 / 3.695	da 436 a 631 / da 498 a 721	[workload-serving-10]
Riparazione	25 passi (il limite del test mantenuto)	3.384 / 3.867	da 540 a 781 / da 617 a 893	[workload-serving-9][workload-serving-10]
Diagnosi	tipica	1.053 / 1.203	circa 210 / 240 (stima)	[workload-serving-10]
Analisi API	tipica (25 endpoint)	4.700 / 5.372	366 / 418	[workload-serving-10]
Analisi API	ogni campo al suo limite	17.076 / 19.515	2.155 / 2.463	[workload-serving-10]

L'istantanea della pagina (fino a 60 elementi interattivi e 3.000 caratteri di testo visibile), la documentazione inclusa e, in una riparazione, il piano stesso dominano i prompt [workload-serving-1][workload-serving-3]. Il budget di risposta predefinito dell'agente (2.048 token) è troppo piccolo per sei suggerimenti API lunghi (2.463 token); lo studio sul carico raccomanda 4.096 [workload-serving-1][workload-serving-10].

1.3 Chiamate per test e per notte

EVENTO	CHIAMATE AL MODELLO	FONTI
Esecuzione con un piano che regge	0	[workload-serving-4]
Codifica di un test da 20 passi	da 8 a 16	[workload-serving-4] [workload-serving-10]
Riparazione di un'esecuzione fallita	1 o 2	[workload-serving-4]
Difetto reale (le riparazioni non lo fanno passare)	1 o 2 riparazioni e 1 diagnosi; all'esecuzione successiva lo stesso piano viene eseguito di nuovo (fino al 16 novembre 2026 l'esecuzione successiva codificava di nuovo il test, da 8 a 16 chiamate)	[workload-serving-4] [workload-serving-67]

Scenari notturni. Le quote di test che si rompono, che falliscono senza modifiche, che falliscono per un difetto reale o che vengono modificati sono ipotesi dello studio sul carico (rotture dell'interfaccia 1%, 3%, 10%; instabili 1,5%, un ordine di grandezza riportato da Google per tutti i suoi tipi di test nel 2016; difetti reali 0,2%, 0,5%, 2%; test modificati 0,5%, 1%, 3%) [workload-serving-9][workload-serving-63][workload-serving-10]. Un difetto reale costa le sue riparazioni e una diagnosi, senza una nuova codifica (la regola del 17 novembre 2026; i valori calcolati il 16 novembre con una nuova codifica sono in "Verifica") [workload-serving-67]. I risultati sono limiti, non previsioni (stima):

SCENARIO	TEST	CHIAMATE	TOKEN DI PROMPT	TOKEN DI RISPOSTA	FONTI
Tipico	100	da 14 a 27	da 51.283 a 116.616	da 2.775 a 8.450	[workload-serving-10]
Tipico	400	da 54 a 106	da 205.130 a 466.463	da 11.098 a 33.801	[workload-serving-10]
Tipico	1.500	da 203 a 398	da 769.238 a 1.749.236	da 41.618 a 126.754	[workload-serving-10]
Notte di rilascio	100	da 40 a 77	da 149.524 a 339.363	da 7.775 a 23.309	[workload-serving-10]
Notte di rilascio	400	da 158 a 308	da 598.096 a 1.357.453	da 31.101 a 93.237	[workload-serving-10]
Notte di rilascio	1.500	da 593 a 1.155	da 2.242.858 a 5.090.447	da 116.629 a 349.637	[workload-serving-10]

Distribuita su una notte di 6 ore, anche la notte di rilascio con 1.500 test fa in media da 104 a 236 token di prompt al secondo [workload-serving-10]. Il vincolo non è il volume, ma il tempo di ogni chiamata.

1.4 Latenza e concorrenza

Un nodo esegue fino a 2, 4 o 8 test alla volta (taglie S, M, L) e accanto a questi un'analisi API, quindi l'endpoint deve servire 3, 5 o 9 richieste concorrenti [workload-serving-4][workload-serving-9]. Il tempo di una chiamata è all'incirca token di prompt / velocità di prefill + token di risposta / velocità di decode + attesa nella coda del server; il timer dell'agente comprende l'attesa [workload-serving-10].

OPERAZIONE (CONTEGGI DI TOKEN IN ITALIANO)	LIMITE RIGIDO NEL CODICE	OBIETTIVO USATO IN QUESTO STUDIO	DECODE NECESSARIO PER RICHIESTA CON UN PREFILL DI 500 / 2.000 TOK/S	FONTI
Giro di codifica, tipico (4.941 in ingresso, 114 in uscita)	60 s	15 s (11 giri e il tempo del browser devono stare in 4 minuti)	22,3 / 9,1 tok/s	[workload-serving-10]
Giro di codifica, documentazione al limite (11.093 in ingresso, 114 in uscita)	60 s	15 s	non raggiungibile / 12,1 tok/s	[workload-serving-10]
Riparazione, 20 passi (3.695 in ingresso, 721 in uscita)	60 s	30 s (l'esecuzione la aspetta)	31,9 / 25,6 tok/s	[workload-serving-10]
Diagnosi, tipica (1.203 in ingresso, 240 in uscita)	30 s	30 s	8,7 / 8,2 tok/s	[workload-serving-10]
Analisi API, tipica (5.372 in ingresso, 418 in uscita)	150 s	150 s	3,0 / 2,8 tok/s	[workload-serving-10]

Le righe della codifica usano 114 token di risposta: i 400 caratteri che la risposta di un giro può occupare quando il server ammette gli spazi, un caso prudente rispetto ai 62-70 token del JSON compatto della sezione 1.2 [workload-serving-10].

In breve: circa 500-2.000 tok/s di lettura del prompt e 25-32 tok/s di generazione per richiesta, con 3-9 richieste in corso. Un server che mette in coda le richieste (Ollama serve per impostazione predefinita una richiesta alla volta per modello) aggiunge l'attesa dentro gli stessi limiti [workload-serving-36]. Dopo un timeout l'agente riprova una volta con un prompt più corto [workload-serving-1].

1.5 Requisiti minimi del modello

REQUISITO	VALORE	FONTI
Output strutturato	JSON conforme a uno JSON Schema, imposto dal server (`response_format` di tipo `json_schema`)	[workload-serving-1] [workload-serving-2]
Parole chiave dello schema	`type`, `properties`, `required`, `additionalProperties: false`, `enum`, `items`, `minItems` / `maxItems`, `pattern` ancorato, `minimum` / `maximum` interi, `maxLength`	[workload-serving-2]
Contesto	almeno 32k token (il massimo teorico della codifica, 33.741 token con la risposta, supera 32.768); 64k raccomandati	[workload-serving-10]
Budget di risposta	4.096 token raccomandati (predefinito 2.048)	[workload-serving-1] [workload-serving-10]
Lingue	istruzioni in inglese; descrizioni, documentazione e testo delle pagine in italiano o in inglese; diagnosi e suggerimenti scritti nella lingua del test	[workload-serving-1] [workload-serving-5]
Ragionamento	disattivato: l'agente si aspetta risposte dirette; i token di ragionamento consumerebbero il budget di risposta e i timeout	[workload-serving-1]
Determinismo	temperatura 0,000001, seed 0	[workload-serving-1]

2. Criteri di selezione

I 120 modelli open-weight esaminati (dati di accompagnamento `dati/models.json`) sono stati filtrati in questo ordine:

01 **Licenza.** Il modello deve poter essere usato commercialmente da una società dell'UE su un nodo nella rete del cliente. Le licenze che escludono l'UE, vietano l'uso commerciale o in produzione, o contengono clausole sul modello di business sono escluse o segnalate per una verifica legale (tabella 3.3).

02 **Memoria.** I pesi alla precisione scelta più nove sessioni da 32k (taglia L) devono entrare nella memoria che resta dopo una riserva (sezione 9). Un modello che contiene meno sessioni serve solo taglie più piccole.

03 **Velocità.** Tempi per chiamata entro gli obiettivi della sezione 1.4 per la taglia, prima come stima (sezione 5), poi misurati sull'hardware del cliente (sezione 7.5).

04 **Output strutturato e uso di strumenti.** L'imposizione dello schema JSON è una funzione del motore di inferenza, non del modello: vLLM impone gli schemi con i backend xgrammar o guidance, llama.cpp con le grammatiche [models-175][models-176]. Le schede dei modelli dichiarano quasi sempre la chiamata di strumenti; le dichiarazioni esplicite sull'output strutturato sono rare (gpt-oss "Structured Outputs", Mistral "JSON output", Nemotron "structured outputs" come obiettivo di addestramento) [models-37] [models-56][models-118].

05 **Italiano.** Dichiarato nella scheda del modello, oppure no. Le schede di Qwen e Gemma dichiarano da 140 a oltre 200 lingue senza nominare l'italiano; GLM elenca solo inglese e cinese [models-3][models-71] [models-79]. Dichiarato non vuol dire misurato (sezione 8.7).

06 **Contesto.** Almeno 32k token nativi.

07 **Ragionamento disattivato.** SparkTest funziona con il ragionamento disattivato, quindi il punteggio rilevante, dove pubblicato, è quello in modalità senza ragionamento: Qwen3.8-27B ottiene 20 sull'indice di Artificial Analysis con lo sforzo "xhigh" e 10 senza ragionamento (indipendente) [models-16].

08 **Supporto del motore.** L'attenzione ibrida (Gated DeltaNet, Mamba-2, finestre scorrevoli), la multi-head latent attention (MLA) e gli indicizzatori sparsi devono essere supportati dalla versione del motore scelta, insieme alla decodifica vincolata dallo schema per il tokenizer del modello [models-175][models-176].

09 **Evidenza indipendente**, dove esiste, con l'harness indicato (sezione 8.1).

Per il vostro reparto IT: applicate gli stessi filtri, nello stesso ordine, a qualunque modello pensiate di portare (sezione 7).

3. I modelli

Selezione: 19 modelli per un singolo nodo (un dispositivo o un sistema compatto) e 9 modelli da data center (server multi-GPU; rilevanti per i clienti che gestiscono un proprio endpoint). I valori sono dichiarati dallo sviluppatore salvo indicazione; "Indice" riporta l'Artificial Analysis Intelligence Index v5.0.1 (AA, indipendente) e l'Epoch Capabilities Index con il suo intervallo al 95% (ECI, indipendente) dove esistono.

3.1 Candidati per un singolo nodo

MODELLO	SVILUPPATORE, USCITA	PARAMETRI TOTALI / ATTIVI	STRATI CON CACHE KV CRESCENTE	CONTESTO NATIVO	LICENZA	ITALIANO NELLA SCHEDA	STRUMENTI / JSON (DICHIARATI)	INDICE (AA; ECI)	FONTI
Qwen3.6-35B-A3B	Qwen (Alibaba), 2026-04	35B / 3B MoE	10 su 40 (gli altri Gated DeltaNet)	256k (262.144)	Apache-2.0	non elencato	strumenti	AA nessun punteggio mostrato; ECI 144,0 (da 142,1 a 145,9)	[models-12][models-14] [benchmarks-15]
Qwen3.8-27B	Qwen, 2026-08	27B denso	16 su 64	256k (262.144)	Apache-2.0	non elencato	strumenti	AA 20 (xhigh), 10 (senza ragionamento); ECI 152,6 (da 150,6 a 154,8)	[models-15][models-16] [benchmarks-15]
Qwen3.5-122B-A10B	Qwen, 2026-02	122B / 10B MoE	12 su 48	256k (262.144)	Apache-2.0	non elencato (201 lingue dichiarate)	strumenti	non trovato	[models-10][models-3]
Qwen3-Coder-Next	Qwen, 2026-01	80B / 3B MoE	12 su 48	256k (262.144)	Apache-2.0	non elencato	strumenti	non trovato	[models-6]
gpt-oss-20b	OpenAI, 2025-08	21B / 3,6B MoE	12 su 24 (12 a finestra scorrevole, finestra 128)	128k (131.072)	Apache-2.0	non indicato	strumenti; Structured Outputs	AA nessun punteggio mostrato; ECI 135,0 (da 132,3 a 137,5)	[models-36][models-37] [models-38][benchmarks-15]
gpt-oss-120b	OpenAI, 2025-08	117B / 5,1B MoE	18 su 36 (18 a finestra scorrevole, finestra 128)	128k (131.072)	Apache-2.0	non indicato	strumenti; Structured Outputs	AA 4 (high); ECI 140,2 (da 138,2 a 142,1)	[models-39][models-37] [models-40][benchmarks-15]
Gemma 4 26B-A4B	Google, 2026-04	25,2B / 3,8B MoE	5 su 30 (25 a finestra scorrevole, finestra 1.024)	256k	Apache-2.0	non elencato (oltre 140 lingue in preaddestramento, oltre 35 supportate)	strumenti	AA nessun punteggio mostrato; ECI 141,6 (da 139,1 a 144,1)	[models-74][models-71] [models-75][benchmarks-15]
Gemma 4 31B	Google, 2026-04	30,7B denso	10 su 60 (50 a finestra scorrevole, finestra 1.024)	256k	Apache-2.0	non elencato (oltre 140 lingue in preaddestramento, oltre 35 supportate)	strumenti	AA 4 (reasoning); ECI 144,3 (da 141,7 a 146,8)	[models-72][models-73] [benchmarks-15]
Muse Glimmer 30B	Meta, 2026-08	29,6B denso (incluso l'encoder visivo)	13 su 52 (39 a finestra scorrevole, finestra 2.048)	128k (131.072) o più	Apache-2.0	non elencato (oltre 100 lingue)	strumenti con schemi precisi	non trovato	[models-48]

MODELLO	SVILUPPATORE, USCITA	PARAMETRI TOTALI / ATTIVI	STRATI CON CACHE KV CRESCENTE	CONTESTO NATIVO	LICENZA	ITALIANO NELLA SCHEDA	STRUMENTI / JSON (DICHIARATI)	INDICE (AA; ECI)	FONTI
Nemotron 3.5 Lightning 30B-A3B	NVIDIA, 2026-08	30B / 3B MoE	6 strati di attenzione (23 Mamba-2)	fino a 1M	OpenMDW-1.1	elencato	strumenti; output strutturati nei dati di addestramento	non trovato nella v5.0.1	[models-118][models-119]
Nemotron 3.5 Bolt 74B-A9B	NVIDIA, 2026-10	74B / 9B MoE	8 strati di attenzione (40 Mamba-2)	fino a 256k	OpenMDW-1.1	elencato	strumenti	non trovato	[models-121][models-119]
Nemotron 3.5 Super VL 120B-A12B	NVIDIA, 2026-10	120B / 12B MoE	8 strati di attenzione (40 Mamba-2)	fino a 1M	OpenMDW-1.1	elencato	strumenti	AA 14	[models-122][models-17]
Mistral Small 4 119B	Mistral AI, 2026-03	119B / 6,5B MoE	36 strati MLA	256k	Apache-2.0	elencato	strumenti; JSON output	AA 4 (reasoning)	[models-56][models-57]
Mistral Small 3.2 24B	Mistral AI, 2025-06	24B denso	40 su 40	128k (131.072)	Apache-2.0	elencato	strumenti	non trovato (BFCL V4: 37,15%, versione ospitata)	[models-52][models-2]
Ministral 3 14B	Mistral AI, 2025-12	14B denso	40 su 40	256k (262.144)	Apache-2.0	elencato	strumenti; JSON output	non trovato	[models-55]
Granite 4.2 30B	IBM, 2026-08	29,3B denso	64 su 64	128k (estensione a 512k)	Apache-2.0	elencato (12 lingue testate)	strumenti	non trovato	[models-113]
Laguna XS 2.1	poolside, 2026-06	33B / 3B MoE	10 su 40 (30 a finestra scorrevole, finestra 512)	256k (262.144)	OpenMDW-1.1	non elencato	strumenti	non trovato	[models-124][models-119]
GLM-4.7-Flash	Z.ai, 2026-01	31B / 3B MoE (dichiarato come "30B-A3B")	47 strati MLA	198k (202.752)	MIT	non supportato (inglese, cinese)	strumenti	non trovato	[models-79]
EuroLLM-22B	progetto UTTER, 2025-12	22,6B denso	54 su 54	32k (32.768)	Apache-2.0	elencato (tutte le lingue ufficiali dell'UE)	non indicato	non trovato	[models-132]

3.2 Modelli da data center

MODELLO	SVILUPPATORE, USCITA	PARAMETRI TOTALI / ATTIVI	CONTESTO NATIVO	LICENZA	ITALIANO NELLA SCHEDA	INDICE (AA; ECI)	FONTI
Qwen3.8-Flash-Next	Qwen, 2026-08	125B + 51B di embedding n-gram + 4B MTP (180B memorizzati) / 6B	256k (262.144)	Qwen Community License 1.0	non elencato	AA 29	[models-21][models-22][models-24]
GLM-5.3-Flash	Z.ai, 2026-08	320B / 18B	1M	MIT	non supportato	AA 28 (pagina del modello); ECI 153,8 (da 152,0 a 155,8)	[models-87][models-88][benchmarks-15]
DeepSeek-V4-Flash-0731	DeepSeek, 2026-07	284B (304B memorizzati) / 13B	1M	MIT	non elencato	AA non trovato; ECI 154,4 (da 152,6 a 156,4)	[models-32][benchmarks-15]
Qwen3.5-397B-A17B	Qwen, 2026-02	397B / 17B	256k (262.144)	Apache-2.0	non elencato	AA non trovato nella v5.0.1; ECI 147,9 (da 145,8 a 150,1)	[models-20][benchmarks-15]

MODELLO	SVILUPPATORE, USCITA	PARAMETRI TOTALI / ATTIVI	CONTESTO NATIVO	LICENZA	ITALIANO NELLA SCHEDA	INDICE (AA; ECI)	FONTI
Nemotron 3 Ultra 550B-A55B	NVIDIA, 2026-06	550B / 55B	fino a 1M	OpenMDW-1.1	elencato	AA non trovato; ECI 147,0 (da 144,8 a 149,2)	[models-123][benchmarks-15]
Mistral Large 3 675B	Mistral AI, 2025-12	675B / 41B	256k	Apache-2.0	elencato	non trovato	[models-64]
DeepSeek-V4.1-Flash-1017	DeepSeek, 2026-10	552B di backbone + 196B di modulo di memoria (763B memorizzati) / 16B in decode	1M	MIT	non elencato	AA 33 (max), 19 (senza ragionamento); ECI 159,4 (da 157,0 a 162,0)	[models-34][models-35][benchmarks-3] [benchmarks-11][benchmarks-15]
GLM-5.3	Z.ai, 2026-08	753B memorizzati / 40B (la scheda dichiara lo stesso modello di base di GLM-5.2 e nessun conteggio; 744B totali e 40B attivi sono dichiarati per GLM-5)	1M	GLM-5.3 License	non supportato	AA 32; ECI 157,1 (da 154,9 a 159,5)	[models-85][models-82][models-86] [benchmarks-15]
Kimi K3.1 Pro	Moonshot AI, 2026-11	2,8T / 104B	1M	Kimi K3.1-Pro License	non elencato	AA 46; ECI 161,4 (da 159,0 a 164,0)	[models-100][models-101][models-17] [benchmarks-15]

Come riferimento, il primo modello proprietario di ciascuna classifica ottiene 65 (AA) e 173,4 (ECI) [benchmarks-4][benchmarks-15].

3.3 Condizioni di licenza che richiedono una decisione

Testi delle licenze letti il 16 novembre 2026; sintesi della nota di accompagnamento, non è un parere legale.

LICENZA	MODELLI	CONDIZIONE COME SCRITTA	EFFETTO PER UN PRODOTTO ON PREMISE	FONTI
Apache-2.0, MIT, OpenMDW-1.1	la maggior parte dei candidati per un singolo nodo	permissive; OpenMDW-1.1 aggiunge una clausola di risoluzione per ritorsione brevettuale e nessun obbligo sugli output	nessuna restrizione sul modello di business trovata	[models-119]
NVIDIA Nemotron Open Model License	Nemotron Nano 9B v2, Nemotron 3 Nano e Super	uso commerciale, nessuna soglia di ricavi; attribuzione; manleva a favore di NVIDIA, anche per gli output	utilizzabile; clausola di manleva da verificare	[models-115]
Qwen Community License 1.0	Qwen3.8-Flash-Next	licenza separata, indipendentemente dai ricavi, per le attività "AI Work Assistant" (programmazione assistita dall'IA o produttività d'ufficio) e Model-as-a-Service	un prodotto che trasforma descrizioni di test in piani di test eseguibili potrebbe essere letto come programmazione assistita dall'IA: serve un parere legale	[models-22]
Modified MIT (Mistral AI)	Mistral Medium 3.5, Devstral 2	nessun diritto se la società del licenziatario ha più di USD 20 milioni di ricavi mensili	se il licenziatario è il cliente, i clienti grandi sono esclusi	[models-59] [models-61]
Licenza Step 5 Preview	Step-5-Preview	licenza separata per i produttori di dispositivi hardware che offrono il modello come funzione del dispositivo	un'offerta in forma di appliance può ricadervi	[models-153]
Muse Model License	Muse Spark Open	"Built with Muse Spark Open" in ogni prodotto che lo usa	obbligo di indicazione in SparkTest	[models-51]
Llama 4 Community License	Llama 4 Scout, Maverick	nessun diritto per persone o società con sede nell'UE sui modelli Llama 4 multimodali; la clausola si trova nella Acceptable Use Policy richiamata dalla licenza e non si applica agli utenti finali di un prodotto che include il modello	escluso per una società italiana che integra il modello nel proprio prodotto	[models-44] [models-45]
Licenza ERNIE 5.5 Flash Base	ERNIE-5.5-Flash-Base	il territorio esclude l'UE, il Regno Unito e la Corea del Sud	escluso	[models-162]
MiniMax M2.7 Non-Commercial	MiniMax-M2.7	uso commerciale solo con autorizzazione scritta preventiva	escluso	[models-146]
Mistral AI Non-Production License	Codestral-22B-v0.1	solo uso non di produzione	escluso	[models-63]

3.4 Note sul panorama

- DeepSeek non ha versioni attuali sotto i 284B di parametri totali (V4-Flash) [models-30]. I modelli aperti di Meta del 2026 sono pubblicati con il nome Muse; non è stata trovata alcuna versione di Llama successiva a Llama 4 [models-48][models-51]. Qwen4 ha un'anteprima ospitata dal 2026-11-11, senza pesi, scheda o licenza pubblicati [models-177][models-178].
- La maggior parte delle versioni del 2026 mantiene una cache KV crescente solo in una parte degli strati (Gated DeltaNet, Mamba-2, finestre scorrevoli); per questo una sessione da 32k costa 0,2 GB su Nemotron 3.5 Lightning contro 5,4 GB su Mistral Small 3.2 (stima, sezione 5.1).
- Holo3-35B-A3B, un fine-tune Apache-2.0 di Qwen3.5-35B-A3B per l'uso del computer, con 3B di parametri attivi, ottiene 82,6 nei risultati ufficiali di OSWorld-Verified (78,2 in una seconda prova negli stessi risultati, 77,8 nella sua scheda) contro 86,0 del primo modello proprietario generalista [benchmarks-34][benchmarks-55]. Mostra che un modello con 3B attivi può usare un'interfaccia utente; SparkTest legge le istantanee del DOM come testo, non le schermate, quindi è un'evidenza indiretta.
- Esistono modelli europei e pensati per l'italiano (Almawave Velvet2-3B, EuroLLM-22B, Apertus 1.5) [models-137][models-132][models-133]. Velvet2-3B non è stato selezionato perché la sua configurazione non è pubblica (repository ad accesso controllato), quindi il suo fabbisogno di memoria non si può ricavare [models-137].

Per il vostro reparto IT: su un singolo dispositivo la scelta pratica è uno dei MoE con 3B attivi della sezione 6.1; per modelli più grandi serve una scheda da 96 GB o un server multi-GPU; consultate la tabella 3.3 prima di scegliere.

4. L'hardware

Specifiche dichiarate dai produttori; i valori misurati sono indicati come tali. I prezzi sono indicativi (osservati): prezzi di strada UE dai comparatori tedeschi o austriaci al 16 novembre 2026, IVA inclusa; i prezzi USA sono di listino senza imposte sulle vendite salvo indicazione; nessuna conversione di valuta. I prezzi delle schede escludono il computer ospite. Tabelle di dettaglio: nota di accompagnamento `fonti/hardware.md` e dati `dati/hardware.json`.

4.1 Classe 1: sistemi compatti a memoria unificata

PIATTAFORMA	MEMORIA	BANDA, DICHIARATA (MISURATA)	CALCOLO IA, DICHIARATO (MISURATO)	POTENZA	PREZZO INDICATIVO	FONTI
NVIDIA GB10, 128 GB (DGX Spark e sistemi dei partner)	128 GB LPDDR5X, condivisi da CPU e GPU	273 GB/s (da 161 a 230 GB/s a riposo, circa da 40 a 105 GB/s durante l'inferenza; comunità)	FP4 1.000 TFLOPS con sparsità (circa da 486 a 511 TFLOPS NVFP4 densi; comunità)	140 W il chip, alimentatore da 240 W	da EUR 5.999 (ASUS Ascent GX10, 1 TB) a EUR 11.383,97 (DGX Spark Founders Edition, 4 TB), prezzo di strada DE; listino USA di DGX Spark USD 6.950 dal 2026-10-02	[hardware-1][hardware-2][hardware-15] [hardware-16][hardware-17][hardware-9] [hardware-8][hardware-4][hardware-5]
NVIDIA GB10, 64 GB (solo sistemi dei partner, dal 2026-10-23)	64 GB LPDDR5X	273 GB/s (stampa)	come sopra	come sopra	da USD 4.999 (prezzo al dettaglio atteso); ASUS Ascent GX10 64 GB da EUR 6.199 (2 offerte), prezzo di strada DE, più del sistema da 128 GB nello stesso giorno	[hardware-4][hardware-5][hardware-10]
AMD Ryzen AI Max+ 395	fino a 128 GB LPDDR5X-8000	256 GB/s (ricavati da bus e velocità; valore del lancio riportato dalla stampa)	GPU FP16 fino a 60 TFLOPS; NPU fino a 50 TOPS	55 W predefiniti, da 45 a 120 W	da EUR 3.899,96 (GMKtec EVO-X2, 128 GB), prezzo di strada DE; Framework Desktop 128 GB USD 1.999 al lancio, USD 3.449 il 2026-09-30	[hardware-18][hardware-22][hardware-21]
AMD Ryzen AI Max+ PRO 495	192 GB LPDDR5X-8533	273 GB/s	GPU FP16 fino a 61 TFLOPS	120 W la piattaforma	da USD 6.799 a 7.449 (listino in preordine, USA, 2026-09-30; primi sistemi consegnati a novembre 2026)	[hardware-19][hardware-21]
NVIDIA Jetson AGX Thor	128 GB LPDDR5X	273 GB/s	FP4 2.070 TFLOPS con sparsità	da 40 a 130 W	kit di sviluppo USD 3.499, portato a USD 5.499 a luglio 2026	[hardware-25][hardware-26]
NVIDIA N1X ("RTX Spark")	fino a 128 GB LPDDR5X-9400	circa 300 GB/s (stampa)	non indicato su base confrontabile	100 W (Dev Box)	Surface RTX Spark Dev Box USD 5.999,99, solo USA	[hardware-27][hardware-28]

Condizioni d'uso: la garanzia di DGX Spark Founders Edition dura un anno, decade in caso di "Enterprise Use" (che la pagina della garanzia lega all'uso in data center o in cluster) e non è trasferibile [hardware-6]; le condizioni di garanzia dei sistemi GB10 dei partner e dei sistemi Ryzen AI Max+ per l'uso presso un cliente non sono state esaminate. Jetson AGX Thor è un modulo per la robotica e N1X una piattaforma PC Windows on Arm; sono elencati per completezza e non vengono confrontati oltre.

4.2 Classe 2: GPU da workstation

SCHEDA	MEMORIA	BANDA	FP8 DENSO / SPARSO (TFLOPS), DICHIARATO	POTENZA DELLA SCHEDA	PREZZO DI STRADA DE/AT, 2026-11-16	FONTI
NVIDIA RTX PRO 6000 Blackwell Workstation Edition	96 GB GDDR7 ECC	1.792 GB/s	1.007,6 / 2.015,2	600 W	da EUR 16.290	[hardware-32][hardware-34][hardware-47]
NVIDIA RTX PRO 6000 Blackwell Max-Q	96 GB GDDR7 ECC	1.792 GB/s	877,9 / 1.755,7	300 W	da EUR 16.489,98	[hardware-33][hardware-34][hardware-48]
NVIDIA RTX PRO 5000 Blackwell, 72 GB	72 GB GDDR7 ECC	1.344 GB/s	non pubblicato (FP4 con sparsità 2.064)	300 W	da EUR 11.999	[hardware-36][hardware-37][hardware-50] [hardware-92]
NVIDIA RTX PRO 5000 Blackwell, 48 GB	48 GB GDDR7 ECC	1.344 GB/s	non pubblicato (FP4 con sparsità 2.064)	300 W	da EUR 8.119	[hardware-36][hardware-37][hardware-49]
NVIDIA RTX PRO 4500 Blackwell	32 GB GDDR7 ECC	896 GB/s	non pubblicato	200 W	da EUR 4.559	[hardware-38][hardware-51]
NVIDIA RTX PRO 4000 Blackwell	24 GB GDDR7 ECC	672 GB/s	non pubblicato (FP4 con sparsità 1.290)	145 W	da EUR 2.780,97	[hardware-39][hardware-52]
NVIDIA GeForce RTX 5090	32 GB GDDR7	1.792 GB/s	838 / 1.676 (accumulo FP16)	575 W	da EUR 7.399,99 (pagina delle varianti MSI; listino UE EUR 2.229, marzo 2025)	[hardware-42][hardware-44][hardware-45]
AMD Radeon AI PRO R9700	32 GB GDDR6	640 GB/s	383 / 766	300 W	da EUR 2.269	[hardware-54][hardware-57]
AMD Radeon PRO W7900	48 GB GDDR6 ECC	864 GB/s	non pubblicato (INT8 123 TOPS)	295 W	nessuna offerta in DE; USD 3.999 al lancio (2023)	[hardware-55][hardware-58]
Intel Arc Pro B70	32 GB GDDR6	608 GB/s	non pubblicato (INT8 367 TOPS)	da 160 a 290 W	da EUR 1.659	[hardware-59][hardware-60][hardware-61]
Intel Arc Pro B60	24 GB GDDR6	456 GB/s	non pubblicato (INT8 197 TOPS)	da 120 a 200 W	da EUR 889	[hardware-59][hardware-62]

Condizioni d'uso: la licenza del driver GeForce afferma che il software GeForce "is not licensed for datacenter deployment" [hardware-43]; AMD afferma che le schede Radeon PRO e Radeon AI PRO sono progettate per workstation, anche montate a rack, e "not designed nor validated for datacenter usage" [hardware-54][hardware-55]. Una NVIDIA DGX Station da scrivania (252 GB HBM3e più 496 GB LPDDR5X) non ha un prezzo di listino; un recensore ha sentito cifre da USD 100.000 a 125.000 [hardware-30][hardware-31].

4.3 Classe 3: acceleratori e server da data center

ACCELERATORE O NODO	MEMORIA PER GPU	BANDA PER GPU	FP8 DENSO / SPARSO (TFLOPS), DICHIARATO	POTENZA	PREZZO INDICATIVO	FONTI
NVIDIA L4 (PCIe)	24 GB GDDR6	300 GB/s	242,5 (ricavato) / 485	72 W	da EUR 3.619, prezzo di strada DE	[hardware-63][hardware-82]
NVIDIA L40S (PCIe)	48 GB GDDR6	864 GB/s	733 / 1.466	350 W	da EUR 9.555, prezzo di strada	[hardware-64][hardware-83]
NVIDIA RTX PRO 6000 Server Edition (PCIe)	96 GB GDDR7	1.597 GB/s	non pubblicato (FP4 "4 PFLOPS", cioè 4.000 TFLOPS, base di sparsità non indicata)	fino a 600 W	da EUR 20.772,07 per scheda, prezzo di strada AT	[hardware-35][hardware-84]
NVIDIA H200 NVL (PCIe)	141 GB HBM3e	4,8 TB/s	1.670,5 (ricavato) / 3.341	fino a 600 W	da EUR 38.900 per scheda, prezzo di strada	[hardware-66][hardware-86]
8 x NVIDIA H100 SXM (HGX)	80 GB HBM3	3,35 TB/s	1.979 (ricavato) / 3.958	fino a 700 W per GPU	USD 195.000 valore equo di mercato, nodo usato (USA, ottobre 2026)	[hardware-65][hardware-87]
8 x NVIDIA H200 SXM (HGX)	141 GB HBM3e	4,8 TB/s	1.979 (ricavato) / 3.958	fino a 700 W per GPU	circa USD 340.000 nuovo, ordine in volume (USA, ottobre 2026)	[hardware-66][hardware-87]

ACCELERATORE O NODO	MEMORIA PER GPU	BANDA PER GPU	FP8 DENSO / SPARSO (TFLOPS), DICHIARATO	POTENZA	PREZZO INDICATIVO	FONTI
8 x NVIDIA B200 (HGX)	180 GB HBM3e (ricavato)	8 TB/s (ricavato)	4.500 / 9.000	circa 14,3 kW per sistema DGX B200	da USD 360.000 a 470.000 nuovo (USA, ottobre 2026)	[hardware-67][hardware-68] [hardware-87][hardware-88]
8 x NVIDIA B300 (HGX)	2,1 TB per 8 GPU (circa 262 GB ciascuna, ricavato)	non elencata	4.500 / 9.000	circa 14 kW per sistema DGX B300	da USD 400.000 a 520.000 (USA, ottobre 2026)	[hardware-68][hardware-69] [hardware-88][hardware-89]
8 x AMD Instinct MI300X	192 GB HBM3	5,3 TB/s	2.614,9 / 5.229,8	750 W di picco	non trovato	[hardware-72]
8 x AMD Instinct MI355X	288 GB HBM3E	8 TB/s	5.000 / 10.100	1.400 W	non trovato	[hardware-75]
Intel Gaudi 3	128 GB HBM2e	3,7 TB/s	1.678 (white paper) o 1.835 (cronache del lancio)	fino a 900 W (OAM)	USD 125.000 per 8 con scheda base (listino 2024, "for modeling only"); prezzo 2026 non trovato	[hardware-77][hardware-91] [hardware-79]

I componenti del 2026 annunciati o in consegna ai fornitori cloud (NVIDIA Rubin con 288 GB HBM4 e 22 TB/s, dati preliminari; AMD Instinct MI455X con 432 GB HBM4 e 23,3 TB/s) e i sistemi a rack (GB300 NVL72 a USD 8,0-8,6 milioni installato) sono fuori dal perimetro di un'installazione SparkTest [hardware-68] [hardware-71][hardware-76][hardware-89].

4.4 Che cosa implicano le tabelle dell'hardware

- La classe 1 offre da 64 a 192 GB in un solo apparecchio a 256-273 GB/s; le schede della tabella 4.2 offrono da 24 a 96 GB per scheda a 456-1.792 GB/s, cioè da 1,7 a 6,6 volte i 273 GB/s di GB10; nella classe 3 i nodi a 8 GPU partono da 80 GB e 3,35 TB/s per GPU, e la configurazione 4 x RTX PRO 6000 Server Edition ha 96 GB e 1.597 GB/s per GPU [hardware-1][hardware-18][hardware-32][hardware-62] [hardware-65][hardware-35].
- La banda limita la velocità di generazione di una richiesta, la memoria limita la dimensione del modello e il contesto concorrente, e il calcolo a bassa precisione limita la lettura del prompt e i batch. Il calcolo dichiarato è confrontabile tra produttori solo sulla stessa base densa o sparsa e con lo stesso formato numerico; NVIDIA afferma che i suoi valori sparsi sono il doppio di quelli densi per FP8, FP16 e INT8 [hardware-63][hardware-67][hardware-68].
- Su un sistema a memoria unificata l'automazione del browser usa la stessa memoria del modello; la banda misurata di GB10 scende a circa 40-105 GB/s con uno strumento che gira accanto all'inferenza [hardware-16][hardware-17]. L'effetto sui tempi delle chiamate di SparkTest non è stato misurato.

5. Ingombro e velocità

Questa sezione confronta le stime del calcolatore con le misure pubbliche. Le stime usano le formule della sezione 9; per le misure pubbliche è indicato chi le ha eseguite. Dati di partenza di ogni stima: `out/tabelle.md` e `out/tabelle.json`.

5.1 Memoria per modello (stima)

Pesi = parametri in memoria x bit per peso / 8, in GB (10^9 byte). Precisioni: BF16 = 16 bit, 8 bit = 8,5 bit (il valore di llama.cpp Q8_0 [method-1]), 4 bit = 4,5 bit (il valore di NVFP4 [method-2]; i file GGUF Q4_K_M sono a 4,89 bit [method-1], circa il 9% in più). Cache KV in BF16 per una sessione da 32k, dalla configurazione di ciascun modello; una cache in FP8 la dimezza.

MODELLO	PARAMETRI IN MEMORIA (B)	ATTIVI (B)	BF16 (GB)	8 BIT (GB)	4 BIT (GB)	KV, UNA SESSIONE DA 32K (GB)	KV, NOVE SESSIONI (GB)
Qwen3.6-35B-A3B	36,0	3,0	71,9	38,2	20,2	0,67	6,0
Qwen3.8-27B	27,8	27,8	55,6	29,5	15,6	2,15	19,3
Qwen3.5-122B-A10B	125,1	10,0	250,2	132,9	70,4	0,81	7,2
Qwen3-Coder-Next	79,7	3,0	159,3	84,6	44,8	0,81	7,2
gpt-oss-20b	20,9	3,6	41,8	22,2	11,8	0,81	7,3
gpt-oss-120b	116,8	5,1	233,7	124,1	65,7	1,21	10,9
Gemma 4 26B-A4B	25,8	3,8	51,6	27,4	14,5	0,88	7,9
Gemma 4 31B	31,3	31,3	62,5	33,2	17,6	3,52	31,7
Muse Glimmer 30B	29,8	29,8	59,6	31,6	16,8	0,52	4,7
Nemotron 3.5 Lightning 30B-A3B	31,6	3,0	63,2	33,6	17,8	0,20	1,8
Nemotron 3.5 Bolt 74B-A9B	74,9	9,0	149,8	79,6	42,1	0,27	2,4
Nemotron 3.5 Super VL 120B-A12B	124,5	12,0	248,9	132,2	70,0	0,27	2,4
Mistral Small 4 119B	119,4	6,5	238,8	126,9	67,2	0,75	6,8
Mistral Small 3.2 24B	24,0	24,0	48,0	25,5	13,5	5,37	48,3
Ministral 3 14B	14,0	14,0	27,9	14,8	7,8	5,37	48,3
Granite 4.2 30B	29,3	29,3	58,6	31,1	16,5	8,59	77,3
Laguna XS 2.1	33,4	3,0	66,9	35,5	18,8	1,41	12,6
GLM-4.7-Flash	31,2	3,0	62,4	33,2	17,6	1,77	16,0
EuroLLM-22B	22,6	22,6	45,3	24,1	12,7	7,25	65,2
Qwen3.8-Flash-Next	180,0	6,0	360,0	191,3	101,3	0,81	7,2
GLM-5.3-Flash	321,3	18,0	642,6	341,4	180,7	non derivata	non derivata
DeepSeek-V4-Flash-0731	304,2	13,0	608,4	323,2	171,1	non derivata	non derivata

MODELLO	PARAMETRI IN MEMORIA (B)	ATTIVI (B)	BF16 (GB)	8 BIT (GB)	4 BIT (GB)	KV, UNA SESSIONE DA 32K (GB)	KV, NOVE SESSIONI (GB)
Qwen3.5-397B-A17B	403,4	17,0	806,8	428,6	226,9	1,01	9,1
Nemotron 3 Ultra 550B-A55B	560,5	55,0	1.121,0	595,6	315,3	0,40	3,6
Mistral Large 3 675B	675,0	41,0	1.350,0	717,2	379,7	2,30	20,7
DeepSeek-V4.1-Flash-1017	763,2	16,0	1.526,4	810,9	429,3	0,03 (dichiarati 890 byte per token)	0,3
GLM-5.3	753,3	40,0	1.506,7	800,4	423,7	2,94	26,5
Kimi K3.1 Pro	2.780,1	104,0	5.560,2	2.953,9	1.563,8	0,91	8,2

I parametri in memoria sono i conteggi safetensors di Hugging Face (comprendono gli encoder visivi e gli strati di predizione multi-token) [models-12][models-15][models-39]; i parametri attivi sono dichiarati, tranne per GLM-5.3, la cui scheda non dà conteggi (i 40B attivi sono dichiarati per GLM-5) [models-85] [models-82]. Gli strati ad attenzione lineare e a spazio degli stati mantengono per ogni sessione uno stato fisso, non incluso. I valori KV del calcolatore coincidono con quelli ricavati nella nota sui modelli per ogni modello dove esistono.

5.2 Matrice di ingombro (stima)

Cella: la prima tra BF16, 8 bit e 4 bit con cui i pesi più **nove** sessioni da 32k (taglia L) entrano accanto alla riserva (16 GB su un sistema a memoria unificata, il 10% della memoria della scheda su una GPU; sezione 9). "4 bit (n)": nessuna precisione contiene nove sessioni; a 4 bit entrano n sessioni concorrenti da 32k. "no": nemmeno una sessione a 4 bit. In questa prova tutte le schede da 32 GB si comportano allo stesso modo, e così tutte quelle da 48 GB e da 24 GB. Una taglia richiede 3 (S), 5 (M) o 9 (L) sessioni concorrenti: una cella come "4 bit (6)" serve solo le taglie S e M.

MODELLO	GB10 64 GB	GB10 128 GB	RYZEN AI MAX+ 395, 128 GB	RYZEN AI MAX+ PRO 495, 192 GB	RTX PRO 6000, 96 GB	RTX PRO 5000, 72 GB	SCHEDA DA 48 GB	SCHEDA DA 32 GB	SCHEDA DA 24 GB
Qwen3.6-35B-A3B	8 bit	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit	4 bit (2)
Qwen3.8-27B	4 bit	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit (6)	4 bit (2)
Qwen3.5-122B-A10B	no	4 bit	4 bit	8 bit	4 bit	no	no	no	no
Qwen3-Coder-Next	4 bit (3)	8 bit	8 bit	BF16	4 bit	4 bit	no	no	no
gpt-oss-20b	8 bit	BF16	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit
gpt-oss-120b	no	4 bit	4 bit	8 bit	4 bit	no	no	no	no
Gemma 4 26B-A4B	8 bit	BF16	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit (8)
Gemma 4 31B	4 bit (8)	BF16	BF16	BF16	8 bit	4 bit	4 bit (7)	4 bit (3)	4 bit (1)
Muse Glimmer 30B	8 bit	BF16	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit
Nemotron 3.5 Lightning 30B-A3B	8 bit	BF16	BF16	BF16	BF16	8 bit	8 bit	4 bit	4 bit
Nemotron 3.5 Bolt 74B-A9B	4 bit	8 bit	8 bit	BF16	8 bit	4 bit	4 bit (4)	no	no
Nemotron 3.5 Super VL 120B-A12B	no	4 bit	4 bit	8 bit	4 bit	no	no	no	no
Mistral Small 4 119B	no	4 bit	4 bit	8 bit	4 bit	no	no	no	no
Mistral Small 3.2 24B	4 bit (6)	BF16	BF16	BF16	8 bit	4 bit	4 bit (5)	4 bit (2)	4 bit (1)

MODELLO	GB10 64 GB	GB10 128 GB	RYZEN AI MAX+ 395, 128 GB	RYZEN AI MAX+ PRO 495, 192 GB	RTX PRO 6000, 96 GB	RTX PRO 5000, 72 GB	SCHEDA DA 48 GB	SCHEDA DA 32 GB	SCHEDA DA 24 GB
Ministral 3 14B	4 bit (7)	BF16	BF16	BF16	BF16	8 bit	4 bit (6)	4 bit (3)	4 bit (2)
Granite 4.2 30B	4 bit (3)	8 bit	8 bit	BF16	4 bit (8)	4 bit (5)	4 bit (3)	4 bit (1)	no
Laguna XS 2.1	4 bit	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit (7)	4 bit (1)
GLM-4.7-Flash	4 bit	BF16	BF16	BF16	BF16	8 bit	4 bit	4 bit (6)	4 bit (2)
EuroLLM-22B	4 bit (4)	BF16	BF16	BF16	4 bit	4 bit (7)	4 bit (4)	4 bit (2)	4 bit (1)

Configurazioni da data center (`*` = cache KV non ricavata, solo pesi):

MODELLO	L40S, 48 GB	RTX PRO 6000 SERVER, 96 GB	H200 NVL, 141 GB	4 X RTX PRO 6000 SERVER, 384 GB	8 X H100, 640 GB	8 X H200, 1.128 GB	8 X B200, 1.440 GB	8 X MI355X, 2.304 GB
Qwen3.8-Flash-Next	no	no	4 bit	8 bit	BF16	BF16	BF16	BF16
GLM-5.3-Flash	no	no	no	8 bit*	8 bit*	BF16*	BF16*	BF16*
DeepSeek-V4-Flash-0731	no	no	no	8 bit*	8 bit*	BF16*	BF16*	BF16*
Qwen3.5-397B-A17B	no	no	no	4 bit	8 bit	BF16	BF16	BF16
Nemotron 3 Ultra 550B-A55B	no	no	no	4 bit	4 bit	8 bit	BF16	BF16
Mistral Large 3 675B	no	no	no	no	4 bit	8 bit	8 bit	BF16
DeepSeek-V4.1-Flash-1017	no	no	no	no	4 bit	8 bit	8 bit	BF16
GLM-5.3	no	no	no	no	4 bit	8 bit	8 bit	BF16
Kimi K3.1 Pro	no	no	no	no	no	no	no	4 bit
gpt-oss-120b	no	4 bit	4 bit	BF16	BF16	BF16	BF16	BF16
Qwen3.6-35B-A3B	4 bit	BF16	BF16	BF16	BF16	BF16	BF16	BF16

Nove sessioni da 32k token sono una prova prudente: i prompt tipici di SparkTest vanno da 4.000 a 11.000 token (sezione 1.2), e i motori con cache KV paginata allocano quello che ogni sessione usa.

5.3 Velocità: stime accanto alle misure pubbliche

La generazione a richiesta singola (decode) è stimata come banda / byte letti per token x un fattore di efficienza tarato su misure pubbliche (sezione 9.3); la lettura del prompt (prefill) come il calcolo efficace misurato sullo stesso hardware / (2 x parametri attivi). Le righe che seguono servono anche alla taratura (sono nel campione), tranne il valore di Ollama (42,1) e quello di vLLM INT4 (32,8), che ne restano fuori; la tabella mostra quindi la dispersione attorno alle mediane, non una validazione indipendente. I valori misurati sono in tok/s, richiesta singola, contesto vuoto o quasi vuoto salvo indicazione. Qui le stime usano i bit del file misurato e la profondità della misura; la sezione 6 usa 4,5 o 8,5 bit per peso alla profondità di una chiamata di codifica, quindi le due tabelle possono differire per la stessa coppia (per esempio gpt-oss-120b su GB10: 55,1 qui, 50 nella sezione 6).

Decode:

MODELLO	HARDWARE	RUNTIME, QUANTIZZAZIONE	MISURATO (CHI)	STIMA	FONTI
gpt-oss-120b	DGX Spark (GB10)	llama.cpp build 7941, MXFP4	58,7 (sviluppatore del runtime)	55,1	[measurements-2]
gpt-oss-120b	DGX Spark (GB10)	vLLM 0.26.1rc1 / Ollama	60,7 / 42,1 (comunità)	55,1	[measurements-12]
gpt-oss-20b	DGX Spark (GB10)	llama.cpp build 7941, MXFP4	83,4 (sviluppatore del runtime)	73,1	[measurements-2]
gpt-oss-120b	Ryzen AI Max+ 395	llama.cpp, ROCm / Vulkan	51,8 / 56,6 (comunità)	51,6	[measurements-20]
Qwen3.6-35B-A3B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL, ROCm / Vulkan	51,5 / 60,4; 51,3 / 60,7 (comunità)	da 73,5 a 75,0	[measurements-20] [measurements-21]
Gemma 4 26B-A4B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	48,5 / 54,7 (comunità)	56,9	[measurements-20]
Qwen3.8-27B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	11,9 (comunità)	11,9	[measurements-21]
Gemma 4 31B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	10,6 / 11,5 (comunità)	11,4	[measurements-20]
Qwen3.5-122B-A10B	Ryzen AI Max+ 395	llama.cpp, UD-Q5_K_XL	20,1 / 22,3 (comunità)	19,4	[measurements-20]
gpt-oss-120b	RTX PRO 6000 Workstation	llama.cpp, MXFP4	196 (sviluppatore del runtime)	179	[measurements-5]
gpt-oss-20b	GeForce RTX 5090	llama.cpp, MXFP4	283 (sviluppatore del runtime)	238	[measurements-5]
Qwen3.6-35B-A3B	Radeon AI PRO R9700	llama.cpp server, Vulkan / ROCm	114,2 / 76,4 (comunità)	91,2	[measurements-21]
Qwen3.8-27B	Radeon AI PRO R9700	llama.cpp server, Vulkan / ROCm	29,3 / 26,6 (comunità)	23,9	[measurements-21]
Qwen3.8-27B	Arc Pro B70	llama.cpp Vulkan / SYCL; vLLM INT4	22,7 / 12,4; 32,8 (comunità)	22,7	[measurements-21]

Prefill:

MODELLO	HARDWARE	MISURATO (CHI)	STIMA	FONTI
gpt-oss-120b	DGX Spark (GB10)	2.444 (llama.cpp build 7941); 1.919 (build 6989); 1.725 (NVIDIA); 1.169 (Ollama)	1.933	[measurements-2][measurements-3] [measurements-9][measurements-11]
gpt-oss-20b	DGX Spark (GB10)	4.506; 3.714; 3.670; 3.224 (stesse fonti)	2.738	[measurements-2][measurements-3] [measurements-9][measurements-11]
Qwen3.6-35B-A3B	Ryzen AI Max+ 395	1.120 / 1.114 (pp512); 1.382 / 1.219 (server, blocchi da 2.048 token)	1.120	[measurements-20][measurements-21]
Qwen3.8-27B	Ryzen AI Max+ 395	409	363	[measurements-21]
gpt-oss-120b	RTX PRO 6000 Workstation	5.518	6.826	[measurements-5]
Qwen3.6-35B-A3B	Radeon AI PRO R9700	3.793 / 3.967	3.793	[measurements-21]
Qwen3.8-27B	Radeon AI PRO R9700	1.022 / 1.080	1.051	[measurements-21]
Qwen3.8-27B	Arc Pro B70	1.742 (vLLM); 1.255 (SYCL); 706 (Vulkan)	1.255	[measurements-21]

L'archivio di [measurements-21] contiene altre prove delle stesse coppie con valori diversi (per esempio vLLM INT4 sulla Arc Pro B70: decode da 31 a 41 tok/s, prefill da 1.740 a 2.211 tok/s); le tabelle mostrano una prova per motore.

Che cosa mostra il confronto:

- La generazione segue la banda: sui sistemi compatti i valori misurati stanno tra il 38% e il 91% del limite di banda, sulle schede tra il 21% e l'82%; i MoE con pochi parametri attivi sulle schede stanno in fondo a questo intervallo (sezione 9.3).
- La stima è alta del 21-46% per Qwen3.6-35B-A3B su Ryzen AI Max+ 395. Per i modelli densi resta entro circa il 20% della maggior parte delle misure; le eccezioni sono effetti del motore (sulla Arc Pro B70 lo stesso modello genera da 12,4 a 32,8 tok/s secondo il motore).
- Il software sposta il risultato quanto l'hardware: il prefill di llama.cpp per gpt-oss-120b su DGX Spark è salito del 27% tra due build [measurements-2][measurements-3]; Ollama ha generato circa il 30% più lentamente di llama.cpp per gpt-oss-120b sulla stessa macchina [measurements-11][measurements-2]; sulla Arc Pro B70 tre motori vanno da 706 a 1.742 tok/s di prefill [measurements-21]; su Ryzen AI Max+ 395 cinque backend software vanno da 234 a 653 tok/s di prefill di gpt-oss-120b a 32k di profondità [measurements-20].
- La profondità del contesto costa dal 22% al 30% della velocità di decode a 32k sui sistemi compatti per gpt-oss-120b, su due build di llama.cpp su DGX Spark e cinque backend su Ryzen AI Max+ 395 [measurements-2][measurements-3][measurements-20].

5.4 Concorrenza

Più richieste contemporanee condividono i pesi letti dalla memoria, quindi la velocità totale di generazione cresce con il numero di richieste, ma ogni richiesta rallenta. Misure: gpt-oss-120b su DGX Spark con llama.cpp è passato da 55 tok/s con una richiesta a 139 tok/s in totale con 8 (17,4 per richiesta) [measurements-2]; con SGLang da 51 a 117 [measurements-7]. In una prova con lo stesso harness, vLLM ha portato gpt-oss-120b da 60,7 tok/s a 116-153 tok/s con 8 richieste concorrenti, mentre Ollama ha dato 42,5 tok/s con 8 richieste contro 42,1 con una sola [measurements-12]. Su una RTX PRO 6000, gpt-oss-120b con SGLang è passato da 208 a 640 tok/s con 8 richieste [measurements-7]. I modelli densi scalano in modo quasi lineare fino a 8 richieste (Gemma 3 27B FP8 con SGLang: da 3,8 a 27,8 tok/s su DGX Spark, da 22,6 a 158,6 su RTX PRO 6000) [measurements-7].

Il calcolatore modella la velocità totale di generazione con c richieste come la velocità a richiesta singola x c^beta, con beta adattato su queste serie: 0,44 per i MoE (da 0,35 a 0,66 su 9 serie) e 0,96 per i modelli densi (da 0,94 a 1,10 su 4 serie). Nella stessa prova vLLM ha portato anche Qwen3-30B-A3B da 77,1 a 313,4 tok/s con 8 richieste (beta 0,67), contro un beta da 0,31 a 0,44 per gpt-oss-120b [measurements-12]: il batching continuo può far scalare alcuni MoE meglio della mediana usata.

5.5 Throughput da data center (misurato, nessuna stima)

HARDWARE	MODELLO	MISURA	VALORE	CHI HA MISURATO	FONTI
H100 / H200 / B200 (per GPU)	gpt-oss-120b, MXFP4, vLLM	tok/s totali per GPU a 50 tok/s per utente	8.973 / 9.319 / 44.489	analista indipendente (InferenceX)	[measurements-32] [measurements-33] [measurements-34]
MI300X / MI355X (per GPU)	gpt-oss-120b, MXFP4, vLLM	idem	8.067 / 38.497	analista indipendente	[measurements-35] [measurements-36]
H100 / H200 / B200 / MI355X (per GPU)	Qwen3.5-397B-A17B, SGLang	idem	1.689 / 1.810 / 9.774 / 8.132	analista indipendente	[measurements-40] [measurements-41] [measurements-42] [measurements-43]
8 x RTX PRO 6000 Server Edition	gpt-oss-120b	MLPerf v6.1 offline, tok/s in uscita per GPU (gli altri risultati per la stessa piattaforma sono più bassi del 2-5%)	circa 1.967	chi ha sottoposto il risultato MLPerf	[measurements-26] [measurements-28]
8 x B200 / 8 x MI355X	gpt-oss-120b	idem	11.436 / 15.227	chi ha sottoposto i risultati MLPerf	[measurements-26]
4 x RTX PRO 6000 Server Edition	DeepSeek-V4-Flash, SGLang	tok/s in uscita, 1 / 16 richieste	39,1 / 349,1	fornitore di servizi	[measurements-51]
2 x DGX Spark	Qwen3.8-Flash-Next, NVFP4	decode a richiesta singola; tempo al primo token, prompt da 32k	29,6 tok/s; 15,9 s	comunità	[measurements-13]

Il fabbisogno di SparkTest alla taglia L è di nove richieste a circa 25-32 tok/s ciascuna (sezione 1.4), una piccola frazione di questi valori. Su un server multi-GPU la questione è di costo, consumo e licenza, non di velocità.

5.6 Misure dell'agente da un capo all'altro

Due prove pubbliche riproducono il lavoro di un agente da un capo all'altro; la decodifica speculativa è attiva in tutte le voci di AgentPerf-Local e nelle voci MLPerf su DGX Spark e Ryzen AI Max+ 395:

PROVA	HARDWARE	MODELLO	RISULTATO	CHI HA MISURATO	FONTI
MLPerf Inference v6.1 Edge Agentic (1.007 turni, contesto da 32k per turno, ragionamento disattivato)	DGX Spark / desktop Ryzen AI Max+ 395	Qwen3.6-27B, NVFP4 con MTP	3,81 s / 7,06 s di media per turno; 29,5 / 17,9 tok/s (dal tempo medio per token in uscita)	chi ha sottoposto il risultato MLPerf (il produttore del motore di inferenza usato)	[measurements-26]
idem	Jetson AGX Thor (risultato sottoposto da NVIDIA) / RTX PRO 4500 Server Edition (32 GB)	Qwen3.6-27B, FP4 / INT4	1,47 s / 2,76 s di media per turno	chi ha sottoposto i risultati MLPerf	[measurements-26]
Artificial Analysis AgentPerf-Local (8 compiti, 168 turni, ragionamento attivo)	DGX Spark / Ryzen AI Halo (Ryzen AI Max+ 395, 128 GB) / RTX 5090	Qwen3.6-35B-A3B, 4 bit	7,3 / 11,7 / 2,0 minuti	analista indipendente	[measurements-25] [measurements-24]
idem	idem	Qwen3.8-27B, 4 bit	17,3 / 34,5 / 4,9 minuti	analista indipendente	[measurements-25] [measurements-24]

Entrambe le prove ordinano i sistemi come il calcolatore. DGX Spark ha completato AgentPerf-Local da 1,6 a 2,0 volte prima del Ryzen AI Halo su questi due modelli; l'articolo riporta da 1,4 a 1,7 volte su tre dei suoi quattro modelli [measurements-25][measurements-24]. Le prove AgentPerf-Local usavano il ragionamento (la pagina dei risultati le indica come "Qwen3.6 35B A3B (Reasoning)" e "Qwen3.8 27B (Xhigh)"), mentre SparkTest lo disattiva; il carico di MLPerf gira con il ragionamento disattivato [measurements-29]. Una sintesi giornalistica di MLPerf v6.1 riporta 19,63 tok/s per la voce Ryzen AI Max+ 395, un valore che non corrisponde ad alcun campo del file ufficiale (token in uscita al secondo sull'intera prova: 10,22) [measurements-29][measurements-26]; questo studio usa i file ufficiali. Entrambe le prove mostrano anche che la decodifica speculativa (MTP) porta un modello denso da 27B su un sistema compatto a circa 29,5 tok/s per richiesta [measurements-26], ben sopra i 6-12 tok/s misurati senza di essa per i modelli densi da 24B a 32B sui sistemi compatti [measurements-11][measurements-20][measurements-21]; il calcolatore non include la decodifica speculativa, il cui guadagno si riduce con la concorrenza [measurements-13].

5.7 Misurato sui compiti di SparkTest

Le prove pubbliche qui sopra ripetono il lavoro di altri agenti. Il 17 novembre 2026 SparkTech ha eseguito il kit di qualificazione dei modelli di SparkTest (sezione 7.5) su due modelli della classe raccomandata nella sezione 6.1: Qwen3.6-35B-A3B (Q4_K_M, 4,74 bit per peso) e gpt-oss-20b (MXFP4, 4,63 bit per peso), serviti da llama.cpp v1.0.0 (build b12359) su un server di prova che non è uno SparkNode: un processore del 2016 (Intel Xeon E5-2630 v4, 10 core, AVX2), 256 GB di memoria DDR4, nessuna GPU, condiviso con altri servizi, il modello su 10 thread alla priorità minima. Il kit manda a ogni modello i prompt dell'agente per otto compiti (tre giri di codifica, un piano intero, due riparazioni, una diagnosi e un'analisi API, in inglese e in italiano) e giudica ogni risposta con i controlli dell'agente. Documento completo, con tutte le risposte conservate: [method-6]. I valori di questa sezione sono misurati da SparkTech, salvo quelli indicati come stime.

MISURA	QWEN3.6-35B-A3B Q4_K_M	GPT-OSS-20B MXFP4	CHE COSA CHIEDE IL KIT
Controlli obbligatori (risposta, formato, validazione dell'agente)	22/23	23/23	tutti
Controlli di qualità (il comportamento atteso)	24/25	25/28	il comportamento che ogni compito chiede
Risposte vincolate da uno schema che superano il controllo dello schema del kit	7/7	7/7	tutte
Riparazione legittima (un pulsante rinominato) accettata dal controllo dello scenario	sì	sì	sì
Riparazione davanti a una pagina con un valore diverso da quello atteso	valore cambiato, riparazione respinta	valore cambiato, riparazione respinta	valore tenuto
Diagnosi in italiano, tre sezioni, prove citate	sì, 118 parole	sì, 119 parole	da 60 a 100 parole, 120 al massimo
Difetti piantati nell'ultimo rilascio trovati dall'analisi API	2 su 2	1 su 2	2 su 2
Chiamate entro i limiti di tempo dell'agente su questo server	1 su 8	1 su 8	8 su 8
Chiamata intera, mediana, su questo server	100,7 s	108,5 s	

Che cosa mostra per questo studio:

- **Le risposte valgono anche per un nodo, i tempi no.** Le risposte dipendono dal modello, dalla sua quantizzazione, dal motore e dal prompt, non dalla velocità della macchina. Su questo processore la lettura del prompt andava a circa 35 tok/s e ha occupato dal 56% al 90% di ogni chiamata; le misure pubbliche degli stessi modelli sull'hardware della classe 1 leggono i prompt da 31 a 124 volte più in fretta (sezione 5.3). La generazione ha raggiunto il 31% (Qwen3.6) e il 52% (gpt-oss-20b) del tetto fissato dai 68,3 GB/s dichiarati, su una sola macchina carica e con una banda dichiarata, non misurata. La velocità di un nodo si misura sul suo hardware (sezione 7.5).
- **Il lavoro decisivo lo fanno i controlli dell'agente.** La grammatica del motore ha fatto seguire a ogni risposta vincolata il suo schema; il validatore dell'agente ha respinto l'unico piano non valido (Qwen3.6 ha messo il testo atteso di due asserzioni nel campo sbagliato); gpt-oss-20b ha compilato un test di accesso negativo con i segnaposto delle credenziali invece dell'utente indicato dal test, cosa che finirebbe in un fallimento visibile. Davanti a una pagina con un valore diverso da quello atteso dal test, entrambi i modelli hanno cambiato il valore atteso e il controllo dello scenario ha respinto le due riparazioni, quindi la regressione è rimasta visibile. Il prompt di riparazione misurato mostrava al modello il titolo del test ma non la sua descrizione, dove il valore è un requisito, e chiedeva di cambiare l'assunzione che fallisce: il risultato misura quel prompt quanto i modelli. Il prompt darà al modello la descrizione e la regola esplicita di non cambiare mai un valore atteso; il controllo dello scenario resta il limite che non dipende dal modello.
- **L'italiano funziona su entrambi i modelli** nella diagnosi e nell'analisi API. Qwen3.6 ha trovato entrambi i difetti piantati nelle statistiche API, gpt-oss-20b uno, e gpt-oss-20b ha affermato un dettaglio sul database che le statistiche non contengono; il kit controlla gli endpoint inventati, non ogni affermazione tecnica.
- **Il ragionamento va spento.** Con l'impostazione predefinita del server, Qwen3.6 ha ragionato fino al tetto di 1.024 token della prova, per 89 s, senza dare una risposta; `chat_template_kwargs` ha spento il ragionamento (risposta in 1,2 s). La versione misurata della richiesta dell'agente non portava quel campo, quindi le misure hanno spento il ragionamento sul server; l'agente lo invia dal 17 novembre 2026 (sezione 7.3).

- **Le ipotesi sui token reggono, salvo per le risposte API.** I prompt di codifica, riparazione e diagnosi hanno avuto da 3,4 a 4,3 caratteri per token (la sezione 1.2 ipotizza 4,0 per l'inglese e 3,5 per l'italiano); il prompt delle statistiche API da 2,6 a 2,8, dal 23% al 35% di token in più del previsto. Le due analisi API sono state di 881 e 700 token di risposta contro i 418 della sezione 1.2, il che porta la stima delle API insights per Qwen3.6 a 4 bit a circa 12 s su GB10 e 16 s su Ryzen AI Max+ 395 (stima), molto sotto il suo limite di 150 s.

Limiti: un campione per compito; una richiesta alla volta, quindi nessuna concorrenza; piani validati, non eseguiti in un browser; solo pesi a 4 bit; ragionamento impostato sul server; i prompt dell'agente sono cambiati dopo la misura (6 degli 8 prompt sono più lunghi di 136 caratteri). È una prima qualificazione di due modelli, non un benchmark, e non dice nulla sulla velocità dell'hardware della sezione 6.

6. Configurazioni confrontate e consigliate

I tempi per chiamata che seguono sono stime per una raffica in cui arrivano insieme tutte le chiamate concorrenti di una taglia, quindi una stima per eccesso; le chiamate isolate sono molto più brevi. Si controllano due casi di prompt: i prompt tipici (giro di codifica da 4.941 token) e la documentazione dell'applicazione al limite di 24.000 caratteri (giro di codifica da 11.093 token); riparazione e diagnosi mantengono i loro prompt tipici in entrambi i casi (sezione 1.4). "Utilizzabile" indica la taglia più grande le cui raffiche rispettano i limiti rigidi dell'agente in entrambi i casi. Il decode è dato alla profondità di una chiamata di codifica tipica (circa 5.000 token) con 4,5 o 8,5 bit per peso; i file GGUF a 4 bit (da 4,9 a 5,2 bit per peso) generano circa dall'8% al 13% più lentamente di questa tabella con llama.cpp (sezione 5.3). "Notte di rilascio" è il numero di test mantenuti il cui carico di una notte di rilascio (limite alto per test) entra in una notte di 6 ore alla concorrenza di ciascuna taglia. Tabelle complete per ogni coppia: `out/tabelle.md`, tabelle 7, 10, 11 e 12.

Una taglia entro i limiti rigidi ma non entro gli obiettivi funziona comunque, più lentamente: i giri di codifica più lunghi di 15 s consumano prima il budget di esplorazione di 4 minuti, quindi l'agente ripiega più spesso su un piano in un colpo solo, e le esecuzioni aspettano più a lungo le riparazioni [workload-serving-4]. Una chiamata che supera il suo limite rigido viene interrotta e riprovata una volta con un prompt più corto [workload-serving-1].

CONFIGURAZIONE	MODELLO, PESI	SESSIONI DA 32K CHE ENTRANO	PREFILL / DECODE ALLA PROFONDITÀ DELLA CODIFICA, TOK/S	CODIFICA: DA SOLA / RAFFICA L (S)	CODIFICA CON DOCUMENTAZIONE AL LIMITE: DA SOLA / RAFFICA M / RAFFICA L (S)	RIPARAZIONE: DA SOLA / S / M / L (S)	OBIETTIVI / LIMITI RIGIDI, PROMPT TIPICI	OBIETTIVI / LIMITI RIGIDI, DOCUMENTAZIONE AL LIMITE	UTILIZZABILE	NOTTE DI RILASCIO ALLA CONCORRENZA DI S / M / L (TEST)
GB10, 128 GB	Qwen3.6-35B-A3B, 4 bit	136	3.286 / 85	2,8 / 18,1	4,8 / 20,4 / 35,2	9,5 / 18,8 / 26,1 / 38,5	M / L	S / L	fino a L	7.960 / 9.105 / 10.463
GB10, 128 GB	Qwen3.6-35B-A3B, 8 bit	109	3.286 / 46	4,0 / 21,9	5,9 / 23,1 / 39,1	16,6 / 31,9 / 43,5 / 62,6	nessuna / M	nessuna / M	fino a M	5.240 / 6.181 / 7.375
GB10, 128 GB	Nemotron 3.5 Lightning, 4 bit	468	3.286 / 89	2,8 / 17,9	4,7 / 20,1 / 34,8	9,2 / 18,3 / 25,4 / 37,6	M / L	S / L	fino a L	8.163 / 9.316 / 10.677
GB10, 128 GB	gpt-oss-120b, 4 bit	38	1.933 / 50	4,8 / 30,8	8,2 / 34,7 / 60,0	16,2 / 32,1 / 44,5 / 65,7	nessuna / M	nessuna / M	fino a M	4.668 / 5.341 / 6.140
GB10, 128 GB	Mistral Small 4, 4 bit	59	1.517 / 40	6,1 / 38,9	10,2 / 43,7 / 75,8	20,2 / 40,0 / 55,6 / 82,1	nessuna / M	nessuna / M	fino a M	3.738 / 4.269 / 4.896
GB10, 128 GB	Qwen3.8-27B, 4 bit	44	701 / 14	15,0 / 72,2	24,0 / 87,9 / 152	55,4 / 68,5 / 80,3 / 103	nessuna / nessuna	nessuna / nessuna	no	2.049 / 2.590 / 3.159
GB10, 64 GB	Qwen3.6-35B-A3B, 4 bit	41	3.286 / 85	2,8 / 18,1	4,8 / 20,4 / 35,2	9,5 / 18,8 / 26,1 / 38,5	M / L	S / L	fino a L	7.960 / 9.105 / 10.463
Ryzen AI Max+ 395, 128 GB	Qwen3.6-35B-A3B, 4 bit	136	1.120 / 80	5,8 / 44,5	11,4 / 53,3 / 94,3	12,2 / 26,3 / 38,3 / 60,0	nessuna / L	nessuna / M	fino a M	4.479 / 4.845 / 5.230
Ryzen AI Max+ 395, 128 GB	Nemotron 3.5 Lightning, 4 bit	468	1.120 / 83	5,8 / 44,4	11,3 / 52,9 / 93,9	11,9 / 25,8 / 37,6 / 59,0	nessuna / L	nessuna / M	fino a M	4.547 / 4.908 / 5.286
Ryzen AI Max+ 395, 128 GB	gpt-oss-120b, 4 bit	38	659 / 47	9,9 / 75,8	19,5 / 90,6 / 160	20,9 / 44,9 / 65,4 / 102	nessuna / S	nessuna / S	fino a S	2.630 / 2.845 / 3.072
Ryzen AI Max+ 395, 128 GB	Qwen3.8-27B, 4 bit	44	363 / 13	22,1 / 132	39,2 / 162 / 285	63,7 / 86,7 / 108 / 151	nessuna / nessuna	nessuna / nessuna	no	1.401 / 1.652 / 1.883
Ryzen AI Max+ PRO 495, 192 GB	Qwen3.5-122B-A10B, 8 bit	53	342 / 14	22,5 / 157	40,6 / 182 / 320	61,6 / 126 / 178 / 269	nessuna / nessuna	nessuna / nessuna	no	1.079 / 1.202 / 1.339
RTX PRO 6000, 96 GB	Qwen3.6-35B-A3B, 8 bit	71	11.604 / 151	1,2 / 6,4	1,7 / 6,7 / 11,3	5,1 / 9,7 / 13,2 / 19,0	L / L	L / L	fino a L	17.410 / 20.608 / 24.701
RTX PRO 6000, 96 GB	gpt-oss-120b, 4 bit	17	6.826 / 162	1,4 / 8,9	2,4 / 10,0 / 17,2	4,9 / 9,7 / 13,4 / 19,8	L / L	M / L	fino a L	15.669 / 18.007 / 20.807
RTX PRO 6000, 96 GB	Qwen3.8-27B, 8 bit	26	3.215 / 40	4,4 / 16,9	6,3 / 20,3 / 34,2	19,0 / 22,2 / 24,9 / 30,0	M / L	S / L	fino a L	7.019 / 9.405 / 12.249
RTX PRO 6000, 96 GB	Mistral Small 4, 4 bit	25	5.356 / 132	1,8 / 11,2	3,0 / 12,6 / 21,7	6,1 / 12,1 / 16,8 / 24,7	L / L	M / L	fino a L	12.553 / 14.400 / 16.604
RTX PRO 6000, 96 GB	Qwen3.5-122B-A10B, 4 bit	19	3.481 / 86	2,7 / 17,3	4,5 / 19,2 / 33,3	9,4 / 18,5 / 25,6 / 37,7	M / L	S / L	fino a L	8.208 / 9.411 / 10.845
RTX PRO 5000, 72 GB	Qwen3.6-35B-A3B, 8 bit	39	5.802 / 113	1,9 / 11,1	3,0 / 12,1 / 20,8	7,0 / 13,6 / 18,7 / 27,2	L / L	M / L	fino a L	11.681 / 13.563 / 15.871
RTX PRO 5000, 72 GB	Qwen3.8-27B, 8 bit	16	1.608 / 30	6,8 / 31,8	10,7 / 38,6 / 66,3	26,1 / 31,9 / 37,0 / 46,9	nessuna / L	nessuna / M	fino a M	4.493 / 5.735 / 7.070
RTX PRO 5000, 48 GB	Qwen3.6-35B-A3B, 4 bit	34	5.802 / 208	1,4 / 9,5	2,5 / 11,0 / 19,2	4,1 / 8,2 / 11,6 / 17,4	L / L	M / L	fino a L	16.971 / 19.062 / 21.447
RTX PRO 5000, 48 GB	Qwen3.8-27B, 4 bit	12	1.608 / 57	5,1 / 29,9	9,0 / 36,7 / 64,4	15,0 / 20,2 / 25,1 / 34,7	S / L	nessuna / M	fino a M	6.081 / 7.210 / 8.258
GeForce RTX 5090, 32 GB	Qwen3.6-35B-A3B, 4 bit	12	9.422 / 277	0,9 / 6,1	1,6 / 7,0 / 12,1	3,0 / 5,9 / 8,3 / 12,2	L / L	L / L	fino a L	24.656 / 28.004 / 31.914
Radeon AI PRO R9700, 32 GB	Qwen3.6-35B-A3B, 4 bit	12	3.793 / 99	2,5 / 15,6	4,2 / 17,6 / 30,5	8,2 / 16,2 / 22,5 / 33,2	M / L	S / L	fino a L	9.234 / 10.558 / 12.126
Radeon AI PRO R9700, 32 GB	Qwen3.8-27B, 4 bit	6	1.051 / 27	8,9 / 47,0	14,9 / 57,4 / 99,8	30,1 / 38,5 / 46,2 / 61,0	nessuna / M	nessuna / M	fino a M	3.453 / 4.246 / 5.037
Arc Pro B70, 32 GB	Qwen3.6-35B-A3B, 4 bit	12	4.529 / 94	2,3 / 13,9	3,7 / 15,4 / 26,4	8,4 / 16,4 / 22,6 / 33,0	M / L	S / L	fino a L	9.516 / 11.015 / 12.838
Arc Pro B70, 32 GB	Qwen3.8-27B, 4 bit	6	1.255 / 26	8,4 / 40,3	13,4 / 49,1 / 84,6	31,0 / 38,3 / 44,9 / 57,4	nessuna / M	nessuna / M	fino a M	3.667 / 4.635 / 5.655
RTX PRO 4000, 24 GB	Nemotron 3.5 Lightning, 4 bit	19	3.626 / 108	2,4 / 15,8	4,1 / 17,9 / 31,2	7,7 / 15,3 / 21,3 / 31,7	M / L	S / L	fino a L	9.572 / 10.863 / 12.368

CONFIGURAZIONE	MODELLO, PESI	SESSIONI DA 32K CHE ENTRANO	PREFILL / DECODE ALLA PROFONDITÀ DELLA CODIFICA, TOK/S	CODIFICA: DA SOLA / RAFFICA L (S)	CODIFICA CON DOCUMENTAZIONE AL LIMITE: DA SOLA / RAFFICA M / RAFFICA L (S)	RIPARAZIONE: DA SOLA / S / M / L (S)	OBIETTIVI / LIMITI RIGIDI, PROMPT TIPICI	OBIETTIVI / LIMITI RIGIDI, DOCUMENTAZIONE AL LIMITE	UTILIZZABILE	NOTTE DI RILASCIO ALLA CONCORRENZA DI S / M / L (TEST)
Arc Pro B60, 24 GB	Nemotron 3.5 Lightning, 4 bit	19	2.215 / 74	3,8 / 25,3	6,6 / 28,9 / 50,4	11,4 / 23,0 / 32,2 / 48,2	S / L	nessuna / L	fino a L	6.210 / 7.006 / 7.922
L40S, 48 GB	Qwen3.6-35B-A3B, 4 bit	34	8.242 / 134	1,5 / 8,3	2,3 / 9,0 / 15,2	5,8 / 11,2 / 15,3 / 22,1	L / L	M / L	fino a L	14.578 / 17.082 / 20.215
H200 NVL, 141 GB	gpt-oss-120b, 4 bit	50	11.048 / 435	0,7 / 4,9	1,3 / 5,7 / 10,0	2,0 / 4,0 / 5,7 / 8,6	L / L	L / L	fino a L	33.953 / 37.936 / 42.427
H200 NVL, 141 GB	Qwen3.8-Flash-Next, 4 bit (licenza: serve un parere legale, tabella 3.3)	31	9.391 / 380	0,8 / 5,8	1,5 / 6,7 / 11,7	2,3 / 4,6 / 6,6 / 9,9	L / L	L / L	fino a L	29.279 / 32.662 / 36.463

Obiettivi: giro di codifica 15 s, riparazione 30 s, diagnosi 30 s; limiti rigidi 60 s, 60 s, 30 s (sezione 1.4). Prefill e decode sono stime (sezione 5.3); le stime di prefill per RTX PRO 5000, RTX PRO 4000, Arc Pro B60, L40S e H200 NVL sono scalate dal calcolo dichiarato perché per queste schede non è stata trovata alcuna misura di prefill (sezione 9.3). La qualità a ciascuna precisione non è stimata (sezione 8.3). Le righe con "no" nella colonna "Utilizzabile" sono riportate per confronto.

6.1 Classe 1: sistema compatto a memoria unificata

- **Hardware:** sistemi NVIDIA GB10 (64 o 128 GB, 273 GB/s dichiarati) e sistemi AMD Ryzen AI Max+ (128 GB a 256 GB/s ricavati, 192 GB a 273 GB/s dichiarati) (sezione 4.1).
- **Modelli (stima):** MoE con circa 3B di parametri attivi a 4 bit: Qwen3.6-35B-A3B, Nemotron 3.5 Lightning 30B-A3B, Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b. Anche GLM-4.7-Flash entra, ma la sua scheda indica solo inglese e cinese [models-79]. Tutti contengono nove sessioni da 32k con margine, anche su 64 GB.
- **Volumi su GB10 (stima):** con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1, entro gli obiettivi fino a M con i prompt tipici (raffica di riparazione a M da 25,4 a 27,8 s contro 30 s) e fino a S con la documentazione al limite (raffica di codifica a M da 20,1 a 20,9 s contro 15 s); utilizzabile fino a L (una raffica di nove riparazioni dura da 37,6 a 40,9 s contro il limite di 60 s). Con Gemma 4 26B-A4B o gpt-oss-20b, entro gli obiettivi fino a S solo con i prompt tipici; utilizzabile fino a L (tabella 12 di `out/tabelle.md`).
- **Volumi su Ryzen AI Max+ 395 (stima):** le chiamate isolate restano entro gli obiettivi (codifica 5,8 s, riparazione 12,2 s), ma nessuna taglia ci resta in raffica, perché la lettura del prompt è circa tre volte più lenta che su GB10 (la raffica di codifica a S dura 15,9 s contro 15 s). Utilizzabile fino a M con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1, fino a S con Gemma 4 26B-A4B o gpt-oss-20b. La taglia L non è consigliata: con i prompt tipici la raffica di riparazione a L dura da 59,0 a 60,0 s contro il limite di 60 s (da 62,5 a 73,8 s per Qwen3.6-35B-A3B alle efficienze di decode più basse della tabella 12), e con la documentazione al limite la raffica di codifica a L dura circa 94 s. Il PRO 495 dà le stesse taglie utilizzabili (tabella 7 di `out/tabelle.md`).
- **Le prove pubbliche da un capo all'altro** confermano l'ordine: DGX Spark ha completato AgentPerf-Local da 1,6 a 2,0 volte prima del Ryzen AI Halo (Ryzen AI Max+ 395, 128 GB) [measurements-25], e un turno di MLPerf Edge Agentic ha richiesto 3,81 s contro 7,06 s [measurements-26] (sezione 5.6).
- **Margine:** la taglia M entro gli obiettivi su GB10 poggia sull'efficienza di decode tarata (0,56), dominata dalle misure di gpt-oss (11 dei 14 punti su GB10). Con l'efficienza misurata per Qwen3.6-35B-A3B su Ryzen AI Max+ 395 (da 0,38 a 0,46), Qwen3.6-35B-A3B e Laguna XS 2.1 scendono a S su GB10, Nemotron 3.5 Lightning a 0,38; la taglia utilizzabile di questi tre modelli non cambia (tabella 12). La taglia M entro gli obiettivi va confermata con la prova di qualifica della sezione 7.5.
- **Precisione:** queste taglie richiedono pesi a 4 bit. A 8 bit GB10 non rispetta gli obiettivi per nessuna taglia ed è utilizzabile fino a M (tabella 6). Il modello a 4 bit deve superare la qualifica della sezione 7.5 su prompt in italiano e lunghi (sezione 8.3).
- **La capacità notturna non è il limite:** una notte di rilascio alla concorrenza della taglia L comprende da 7.900 a 10.700 test mantenuti su GB10 e da 4.000 a 5.300 su Ryzen AI Max+ 395 con i cinque modelli sopra (stima).
- **Che cosa non fa:** la taglia L entro gli obiettivi; i modelli densi da 24B a 32B entro gli obiettivi (da 6 a 12 tok/s di generazione misurati, 13-14 stimati; una riparazione da sola richiede circa 55-64 s, stima), a meno che la decodifica speculativa non sia qualificata sul motore scelto [measurements-26]; i MoE da 117B a 125B entro gli obiettivi: gpt-oss-120b e Mistral Small 4 sono utilizzabili fino a M su GB10 (raffica di riparazione a M 44,5 e 55,6 s), e su Ryzen AI Max+ 395 solo gpt-oss-120b, fino a S (la sua raffica di riparazione a M dura 65,4 s, oltre il limite di 60 s); il sistema da 192 GB consente pesi a 8 bit per i modelli della classe da 120B, ma non la velocità per usarli con le latenze di SparkTest; i modelli da data center. La garanzia di DGX Spark Founders Edition esclude l'"Enterprise Use" [hardware-6].

6.2 Classe 2: singola GPU da workstation

- **Hardware:** una scheda in una workstation o in una workstation a rack: RTX PRO 6000 (96 GB), RTX PRO 5000 (48 o 72 GB), schede da 32 GB (RTX PRO 4500, Radeon AI PRO R9700, Arc Pro B70; GeForce RTX 5090 con la restrizione di licenza indicata sotto) (sezione 4.2). Le schede da 24 GB (RTX PRO 4000, Arc Pro B60) sono un'opzione più limitata, descritta in fondo ai volumi.
- **Modelli (stima):** con 96 GB, tutti i candidati per un singolo nodo della tabella 3.1: MoE con 3B attivi a 8 bit o in BF16, densi da 27B a 31B a 8 bit, MoE da 117B a 125B a 4 bit (Granite 4.2 30B, la cui cache KV ad attenzione completa occupa 8,6 GB per sessione da 32k, contiene otto sessioni a 4 bit). Con 72 GB, MoE con 3B attivi a 8 bit o in BF16 e densi da 27B a 31B a 4 o 8 bit. Con 48 GB, MoE con 3B attivi a 4 o 8 bit (nove sessioni a 8 bit per Nemotron 3.5 Lightning, Gemma 4 26B-A4B e gpt-oss-20b) e densi da 27B a 30B a 4 bit; Gemma 4 31B, con la sua cache KV più grande, vi contiene sette sessioni da 32k a 4 bit (taglia M). Con 32 GB, MoE con 3B attivi a 4 bit (Laguna XS 2.1 e GLM-4.7-Flash contengono solo sette e sei sessioni: al massimo la taglia M); un denso da 27B entra a 4 bit ma contiene solo sei sessioni da 32k (taglia M). Con 24 GB, gpt-oss-20b e Nemotron 3.5 Lightning contengono nove sessioni a 4 bit e Gemma 4 26B-A4B otto; gli altri modelli con 3B attivi due o meno.

- **Volumi (stima):**
  - MoE con 3B attivi: sulla RTX PRO 6000, entro gli obiettivi fino a L con entrambi i casi di prompt (Qwen3.6-35B-A3B a 8 bit); sulla RTX PRO 5000 (48 o 72 GB), fino a L con i prompt tipici e M con la documentazione al limite; sulla GeForce RTX 5090, fino a L in entrambi i casi (Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, gpt-oss-20b); sulla R9700 e sulla Arc Pro B70, fino a M con i prompt tipici e S con la documentazione al limite, utilizzabile fino a L (Gemma 4 26B-A4B: S con i prompt tipici; Laguna XS 2.1: utilizzabile fino a M); sulla W7900, il cui prefill è stimato solo dal calcolo dichiarato, fino a S con i prompt tipici, utilizzabile fino a M. La RTX PRO 4500 non è stimata (nessun valore di calcolo pubblicato).
  - gpt-oss-120b e Mistral Small 4 a 4 bit su 96 GB: entro gli obiettivi fino a L con i prompt tipici e M con la documentazione al limite. Un denso da 27B a 8 bit su 96 GB: entro gli obiettivi fino a M con i prompt tipici e S con la documentazione al limite, utilizzabile fino a L; sulle schede da 32 a 72 GB un denso da 27B è utilizzabile fino a M.
  - Schede da 24 GB con Nemotron 3.5 Lightning a 4 bit: la RTX PRO 4000 entro gli obiettivi fino a M con i prompt tipici e S con la documentazione al limite, la Arc Pro B60 fino a S solo con i prompt tipici; entrambe utilizzabili fino a L (con gpt-oss-20b la Arc Pro B60 è utilizzabile fino a M).
  - Capacità notturna: da 3.500 a 31.900 test secondo scheda, modello e concorrenza.
- **Che cosa non fa:** i MoE da 117B a 125B sotto i 96 GB; i modelli da data center; il software GeForce "is not licensed for datacenter deployment" [hardware-43]; le schede Radeon PRO e Radeon AI PRO sono "not designed nor validated for datacenter usage" [hardware-54][hardware-55]. Prezzi delle schede al 16 novembre 2026, computer ospite escluso: da EUR 1.659 (Arc Pro B70, 32 GB) a EUR 16.489,98 (RTX PRO 6000 Max-Q); schede da 24 GB da EUR 889 (Arc Pro B60) (osservati) [hardware-61][hardware-48] [hardware-62].

6.3 Classe 3: server multi-GPU

- **Hardware:** 4 x RTX PRO 6000 Server Edition (384 GB), nodi a 8 GPU con H100, H200, B200 o MI355X (da 640 GB a 2,3 TB) (sezione 4.3). Le singole schede da data center (L40S, H200 NVL) si comportano per SparkTest come la classe 2.
- **Modelli:** modelli da data center per i clienti che gestiscono già una piattaforma di modelli o vogliono un modello più grande: Qwen3.5-397B-A17B (4 bit su 4 x RTX PRO 6000 Server, 8 bit su 8 x H100), GLM-5.3-Flash e DeepSeek-V4-Flash (8 bit su 4 x RTX PRO 6000 Server, solo pesi), Nemotron 3 Ultra, Mistral Large 3, DeepSeek-V4.1-Flash; GLM-5.3 e Kimi K3.1 Pro con licenze proprie (tra le configurazioni confrontate, Kimi K3.1 Pro entra solo in 8 x MI355X a 4 bit); Qwen3.8-Flash-Next solo dopo un parere legale sulla sua licenza (tabella 3.3).
- **Volumi:** non sono un vincolo; il throughput misurato è di ordini di grandezza superiore al carico di SparkTest alla taglia L (sezione 5.5). Un server può servire più SparkNode, purché l'endpoint sia raggiungibile solo dall'interno della rete del cliente (sezione 7.4).
- **Che cosa non fa:** non è un'appliance; i nodi a 8 GPU costano da USD 195.000 (H100 usato) a USD 470.000 (B200 nuovo) negli USA a ottobre 2026, e un sistema DGX B200 o DGX B300 assorbe fino a circa 14 kW (dichiarati) [hardware-87][hardware-88][hardware-68][hardware-69]; il vantaggio dei loro modelli sul compito di SparkTest non è misurato in modo indipendente; la cache KV di alcune architetture del 2026 non si è potuta ricavare (GLM-5.3-Flash, DeepSeek-V4-Flash).

6.4 Scelta per taglia

Le stesse stime lette per taglia. "MoE con 3B attivi" indica Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, Laguna XS 2.1, Gemma 4 26B-A4B o gpt-oss-20b; le eccezioni sono indicate.

TAGLIA	ENTRO GLI OBIETTIVI CON ENTRAMBI I CASI DI PROMPT	ENTRO GLI OBIETTIVI SOLO CON I PROMPT TIPICI	SOLO ENTRO I LIMITI RIGIDI (FUNZIONA, PIÙ LENTAMENTE)
S (2 esecuzioni parallele)	GB10 con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1 a 4 bit; Radeon AI PRO R9700 e Arc Pro B70 con MoE con 3B attivi a 4 bit; RTX PRO 4000 con Nemotron 3.5 Lightning o gpt-oss-20b; RTX PRO 5000, RTX PRO 6000 e GeForce RTX 5090 con MoE con 3B attivi	GB10 con Gemma 4 26B-A4B o gpt-oss-20b; Arc Pro B60 con Nemotron 3.5 Lightning	Ryzen AI Max+ 395 o PRO 495 con MoE con 3B attivi a 4 bit
M (4 esecuzioni parallele)	RTX PRO 5000 (48 o 72 GB), RTX PRO 6000 o GeForce RTX 5090 con MoE con 3B attivi	GB10 con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1 (margine ridotto, sezione 6.1); Radeon AI PRO R9700 e Arc Pro B70 con MoE con 3B attivi tranne Gemma 4 26B-A4B; RTX PRO 4000 con Nemotron 3.5 Lightning o gpt-oss-20b	GB10 con Gemma 4 26B-A4B, gpt-oss-20b, gpt-oss-120b o Mistral Small 4; Ryzen AI Max+ 395 o PRO 495 con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o Laguna XS 2.1; R9700 e Arc Pro B70 con Gemma 4 26B-A4B; Arc Pro B60 con Nemotron 3.5 Lightning o gpt-oss-20b
L (8 esecuzioni parallele)	RTX PRO 6000 con MoE con 3B attivi a 8 bit; GeForce RTX 5090 con Qwen3.6-35B-A3B, Nemotron 3.5 Lightning o gpt-oss-20b	RTX PRO 5000 (48 o 72 GB) con MoE con 3B attivi; RTX PRO 6000 con gpt-oss-120b o Mistral Small 4; GeForce RTX 5090 con Gemma 4 26B-A4B	GB10 con MoE con 3B attivi a 4 bit; Radeon AI PRO R9700 e Arc Pro B70 con MoE con 3B attivi tranne Laguna XS 2.1; RTX PRO 4000 con Nemotron 3.5 Lightning o gpt-oss-20b; Arc Pro B60 con Nemotron 3.5 Lightning; RTX PRO 6000 con un denso da 27B a 8 bit

Tutte le voci sono stime (sezione 6) e vanno confermate con la prova di qualifica della sezione 7.5; i sistemi Ryzen AI Max+ non sono utilizzabili alla taglia L. La GeForce RTX 5090 ha la restrizione di licenza della sezione 4.2; la Radeon PRO W7900 è trattata nella sezione 6.2.

7. Portare il proprio modello

7.1 Requisiti

REQUISITO	CHE COSA SERVE	PERCHÉ	FONTI
Endpoint	`POST {base}/v1/chat/completions` compatibile OpenAI, senza streaming, raggiungibile dallo SparkNode all'interno della rete del cliente	l'agente invia una richiesta per operazione	[workload-serving-1]
Output strutturato	`response_format` di tipo `json_schema`, imposto dal motore, con le parole chiave della sezione 1.5	codifica, riparazione e analisi API interpretano la risposta come JSON	[workload-serving-1] [workload-serving-2]
Contesto	almeno 32k token per richiesta, 64k raccomandati; dimensionato per slot concorrente dove il server condivide un unico pool	prompt da 4k a 11k token, massimo teorico sopra 32k	[workload-serving-10] [workload-serving-29]
Budget di risposta	`LLM_MAX_TOKENS` portato a 4.096	sei suggerimenti API lunghi richiedono circa 2.460 token	[workload-serving-10]
Ragionamento	disattivato per le chiamate di SparkTest (`reasoning_effort: "none"`, un'opzione del template, un flag del server o un modello senza ragionamento); l'agente lo chiede (7.2), il kit mostra se il server lo rispetta (5.7)	i token di ragionamento consumano il budget e i timeout	[workload-serving-16] [workload-serving-24] [workload-serving-29] [workload-serving-67]
Concorrenza	tante richieste simultanee quante sono le esecuzioni parallele della taglia più una (3, 5 o 9), senza code oltre i timeout	sezioni 1.4 e 5.4	[workload-serving-4] [workload-serving-9]
Latenza	p50 e p95 per operazione entro gli obiettivi della sezione 1.4 con quella concorrenza	i timer dell'agente comprendono la coda	[workload-serving-10]
Perimetro dei dati	endpoint dentro il perimetro del cliente, legato a un indirizzo interno, con chiave e TLS	i prompt contengono descrizioni dei test, documentazione, testo ed etichette delle pagine, messaggi di errore e URL delle API (segreti mascherati)	[workload-serving-1] [workload-serving-3]

7.2 Collegare l'endpoint

L'agente passa all'endpoint del cliente quando `LLM_URL` è impostata. Impostazioni lette dal codice [workload-serving-1]:

VARIABILE	SIGNIFICATO	PREDEFINITO
`LLM_URL`	URL di base; l'agente aggiunge `/v1/chat/completions`	non impostata
`LLM_MODEL`	nome del modello servito (vLLM `--served-model-name`, tag di Ollama)	`default_model`
`LLM_API_KEY`	inviata come `Authorization: Bearer` quando impostata	vuota
`LLM_MAX_TOKENS`	budget di risposta	2.048 (portare a 4.096)
`LLM_TEMPERATURE`	temperatura di campionamento (il seed 0 viene sempre inviato)	0,000001
`LLM_TIMEOUT_MS`	timeout delle chiamate di codifica e di riparazione	60.000
`LLM_EFFORT`	con `low` la richiesta chiede di non ragionare: `enable_thinking: false` al primo livello e, dal 17 novembre 2026, in `chat_template_kwargs`, più `reasoning_effort: "low"`	`low`

La richiesta contiene `model`, un solo messaggio utente, `max_tokens`, `temperature`, `seed`, i campi sul ragionamento indicati sopra e, per codifica, riparazione e analisi API, `response_format: {"type": "json_schema", "json_schema": {"name": "sparktest_answer", "schema": ...}}` (il nome dal 17 novembre 2026); la risposta viene letta da `choices[0].message.content` [workload-serving-1][workload-serving-67]. Per un'autorità di certificazione interna, Node.js legge `NODE_EXTRA_CA_CERTS` o `NODE_USE_SYSTEM_CA=1` [workload-serving-65].

7.3 Software di serving: compatibilità al 16 e al 17 novembre 2026

Tutte le celle sono dichiarazioni del produttore (documentazione, note di rilascio, codice sorgente) lette dallo studio sul carico il 16 novembre 2026, applicate alla richiesta dell'agente com'è dal 17 novembre 2026; finora solo llama.cpp ha servito le richieste di SparkTest (sezione 5.7).

SERVER (VERSIONE VERIFICATA)	LICENZA	ACCETTA LA RICHIESTA DELL'AGENTE (DAL 17 NOVEMBRE 2026)	RAGIONAMENTO DISATTIVATO	CONCORRENZA	NOTE	FONTI
vLLM 0.33.0	Apache-2.0	sì, stando al codice sorgente: l'agente ora invia `json_schema.name`, che è obbligatorio (fino al 16 novembre: no)	`reasoning_effort: "none"` o `chat_template_kwargs` (l'agente invia il secondo); il campo `enable_thinking` al primo livello viene ignorato	batching continuo	statistiche d'uso attive per impostazione predefinita (si disattivano con `VLLM_NO_USAGE_STATS` o `DO_NOT_TRACK`); immagini di rilascio compilate per CUDA 13.0, quindi driver NVIDIA 580 o successivo	[workload-serving-16][workload-serving-20][workload-serving-12][workload-serving-22]
SGLang 0.5.23	Apache-2.0	sì, stando al codice sorgente (fino al 16 novembre: no, `name` obbligatorio)	`reasoning_effort: "none"`, che imposta l'interruttore del template che l'agente ora invia in `chat_template_kwargs`	non verificata	cache dei prefissi RadixAttention attiva per impostazione predefinita	[workload-serving-24][workload-serving-26]
llama.cpp server b12445	MIT	sì (viene letto solo `json_schema.schema`; misurato con b12359, sezione 5.7)	`chat_template_kwargs` (inviato dall'agente; misurato su Qwen3.6), `reasoning_effort: "none"` o `--reasoning-budget 0`	slot paralleli, batching continuo	schema convertito in grammatica; le parole chiave non supportate vengono saltate in silenzio; `--offline`; si lega a 127.0.0.1 per impostazione predefinita	[workload-serving-29][workload-serving-30][workload-serving-31]
Ollama 0.40.16	MIT	sì	`reasoning_effort: "none"`; l'agente invia `"low"`, quindi usate un modello senza ragionamento o verificate con il kit che non ne torni	una richiesta per modello per impostazione predefinita (`OLLAMA_NUM_PARALLEL`)	contesto predefinito di 4k sotto i 24 GiB di memoria GPU; nessuna autenticazione sull'API locale	[workload-serving-33][workload-serving-35][workload-serving-36][workload-serving-37][workload-serving-40]
LM Studio 0.4.x	applicazione con proprie condizioni d'uso (gratuita per il lavoro); CLI e SDK MIT	non documentato	non documentato	batching continuo nel motore llama.cpp	il serving funziona offline	[workload-serving-41][workload-serving-43][workload-serving-45][workload-serving-46]
NVIDIA NIM 2.1.6	NVIDIA AI Enterprise in produzione: listino USD 4.500 per GPU all'anno	segue il suo motore (vLLM 0.32.0 o SGLang 0.5.22), che richiede `name`: inviato dal 17 novembre 2026	come il suo motore	come il suo motore	installazione air-gap documentata	[workload-serving-50][workload-serving-52][workload-serving-54]
OpenVINO Model Server 2026.4.1	Apache-2.0	il percorso è `/v3/chat/completions`	non verificato	non verificata	se non riesce a elaborare il formato, procede senza, e senza errore	[workload-serving-57]
Hugging Face TGI 3.3.7	Apache-2.0	no: forma di `response_format` diversa	non verificato	non verificata	repository archiviato; modalità di manutenzione	[workload-serving-48][workload-serving-49]

Le modifiche che rendono portabile la richiesta dell'agente sono piccole e riguardano l'agente di SparkTest, non l'endpoint del cliente. Due sono state fatte il 17 novembre 2026: la richiesta invia `json_schema.name` e chiede di non ragionare con `chat_template_kwargs` e `reasoning_effort: "low"` oltre al campo al primo livello (`"low"` è il livello più basso dei modelli come gpt-oss, che ragionano sempre; `"none"` non viene inviato) [workload-serving-67]. Restano aperte: `strict`, un percorso dell'endpoint configurabile, un budget di risposta predefinito di 4.096 token e la registrazione dei conteggi di token di `usage` [workload-serving-16][workload-serving-24][workload-serving-57]. Finché il percorso non è configurabile, OpenVINO Model Server richiede una riscrittura in un reverse proxy (sezione 7.4).

I motori di imposizione dello schema dietro questi server (XGrammar, llguidance, Outlines, grammatiche di llama.cpp) sono stati confrontati in modo indipendente su JSONSchemaBench, con modelli Llama piccoli (da 1B a 8B di parametri): sugli schemi semplici di chiamata di funzioni la loro copertura era da 0,93 a 0,96, sugli schemi difficili raccolti da GitHub da 0,03 a 0,41 [workload-serving-58]. I tre schemi di SparkTest sono piccoli (da 494 a 945 caratteri) e usano parole chiave comuni, quindi il rischio di copertura è basso; la conformità va comunque dimostrata per ogni modello [workload-serving-17][workload-serving-30].

7.4 Lista di controllo per il reparto IT del cliente

01 Endpoint: `POST {base}/v1/chat/completions`, senza streaming, raggiungibile solo dallo SparkNode; altri percorsi (`/v3`) richiedono una riscrittura in un reverse proxy [workload-serving-57].

02 `LLM_MODEL` uguale al nome del modello servito [workload-serving-19].

03 Imposizione dello schema attiva, con un motore che copra le parole chiave della sezione 1.5; dimostrate che è imposta, non solo accettata, con una prova volutamente severa (un enum, un array con limiti), perché alcuni server saltano in silenzio parole chiave o formati non supportati [workload-serving-30] [workload-serving-57].

04 JSON compatto dove il server lo consente (vLLM `disable_any_whitespace`), per risparmiare token di risposta [workload-serving-17].

05 Contesto di 32k per richiesta, dimensionato per slot dove il pool è condiviso; per Ollama impostate `OLLAMA_CONTEXT_LENGTH` e `OLLAMA_NUM_PARALLEL` [workload-serving-29][workload-serving-35] [workload-serving-36].

06 Ragionamento disattivato per le chiamate di SparkTest (7.3), confermato con il kit di qualificazione, che riporta qualsiasi ragionamento restituito dal server (sezione 5.7).

07 Concorrenza della taglia senza code oltre i timeout; misurate p50 e p95 per operazione.

08 Chiave e TLS sull'endpoint (i flag `--api-key` e TLS esistono in vLLM, SGLang e llama.cpp; l'API locale di Ollama richiede un reverse proxy) [workload-serving-19][workload-serving-26][workload-serving-29] [workload-serving-37].

09 Nessuna chiamata in uscita dallo stack di serving: `VLLM_NO_USAGE_STATS=1` o `DO_NOT_TRACK=1`, `HF_HUB_OFFLINE=1`, llama.cpp `--offline`, NIM in modalità air-gap [workload-serving-20][workload-serving-61][workload-serving-29][workload-serving-52].

10 Registrazione delle richieste del software di serving verificata, dato che i prompt contengono i dati elencati in 7.1.

11 Licenza del software di serving e del modello verificata (sezioni 3.3 e 7.3).

12 Prova di qualifica superata (7.5).

7.5 Qualifica con un insieme di test

I benchmark pubblici non misurano il compito di SparkTest (sezione 8.1), quindi ogni modello ed endpoint viene qualificato sui casi di SparkTest prima dell'uso, con la validazione e il controllo dello scenario dell'agente attivi come in produzione [workload-serving-1][workload-serving-2]. SparkTech qualifica allo stesso modo ogni SparkNode prima dell'uso.

01 **Insieme di test:** test mantenuti scritti in italiano e in inglese su applicazioni di riferimento, con prompt con la documentazione al limite, modifiche dell'interfaccia registrate per le riparazioni e difetti reali inseriti apposta (una riparazione non deve trasformare un difetto reale in un test superato).

02 **Esecuzioni:** codificare ogni test dalla sua descrizione; eseguire la suite su una versione modificata dell'applicazione; ripetere alla concorrenza della taglia.

03 **Misure:** validità del JSON; quota di test codificati al primo tentativo; quota di rotture riparate e accettate dal controllo dello scenario; riparazioni che hanno nascosto un difetto reale; diagnosi scritta nella lingua del test; tempo p50 e p95 per operazione; token di prompt e di risposta dal campo `usage`.

04 **Confronto:** la stessa prova su una configurazione di riferimento. Le soglie di accettazione si concordano con il cliente; le soglie di tempo seguono la sezione 1.4.

Il kit di qualificazione di SparkTech (`agent/bin/qualify.js`) copre le prime misure di questo elenco su otto compiti fissi costruiti con i prompt e i controlli dell'agente, contro qualunque endpoint compatibile OpenAI, una richiesta alla volta; la sua prima esecuzione, su due modelli open-weight, è la sezione 5.7. La concorrenza e i test del cliente si misurano sul nodo.

8. Rischi e limiti

8.1 Benchmark dichiarati e chi li ha misurati

- I dati del produttore e quelli indipendenti per lo stesso modello differiscono molto: gpt-oss-120b ottiene 62,4 su SWE-bench Verified nella sua scheda (sforzo high, sottoinsieme di 477 compiti) contro il 33,60% di Vals AI e 26,0 sulla classifica bash-only [benchmarks-49][benchmarks-18][benchmarks-23].
- Le prove di due concorrenti sullo stesso modello differiscono di 11,7 punti (Qwen3.6-27B su OSWorld-Verified) e di 14,6 punti (Gemma 4 31B su SWE-bench Verified); due harness sullo stesso benchmark di 11-13 punti [benchmarks-42][benchmarks-48][benchmarks-45][benchmarks-18][benchmarks-54].
- Un modello che non segue il protocollo di un harness può ottenere un punteggio vicino a zero: Nemotron 3.5 Super VL ottiene lo 0,00% su Terminal-Bench 4.0 con un harness minimale, mentre il suo sviluppatore riporta 60,67 su Terminal-Bench 2.1 con un altro harness [benchmarks-17][benchmarks-54].
- I dati indipendenti sui modelli per un singolo nodo sono scarsi: il Berkeley Function Calling Leaderboard V4 (ultimo aggiornamento 2026-04-12) non contiene alcun modello open-weight del 2026 [benchmarks-32]; non è stato trovato alcun dato indipendente di Terminal-Bench 4.x, MCP Atlas o tau3 per Qwen3.8-27B, Gemma 4 31B, Muse Glimmer 30B, Qwen3.6-35B-A3B o Nemotron 3.5 Lightning.

Per il vostro reparto IT: non scegliete un modello dalle classifiche pubbliche; la prova che conta è la qualifica della sezione 7.5.

8.2 Saturazione e contaminazione

Su SWE-bench Verified, i modelli open-weight usciti da luglio 2026 ottengono dall'86,0% (Qwen3.8-27B) al 96,4% (DeepSeek-V4-Pro-0813) presso Vals AI, che non lo esegue più sui nuovi modelli; OpenAI ha smesso di riportarlo a febbraio 2026 dopo aver trovato test difettosi nel 59,4% di 138 compiti verificati e indizi che i modelli avessero visto dettagli dei compiti durante l'addestramento [benchmarks-18][benchmarks-24][benchmarks-25]. I problemi di LiveCodeBench v6 sono anteriori ai dati di addestramento dei modelli del 2026 [benchmarks-54]. Punteggi vicini al 90% su questi insiemi dicono poco delle differenze che contano per SparkTest.

8.3 Perdita di qualità con la quantizzazione

- Uno studio accademico indipendente ha trovato che la quantizzazione a 8 bit preserva l'accuratezza (circa lo 0,8% in meno), mentre i metodi a 4 bit perdono fino al 59% sui compiti a contesto lungo, con perdite che dipendono da metodo, modello e compito e che crescono quando l'input non è in inglese [method-4]. I prompt di SparkTest sono lunghi e spesso in italiano.
- Una prova indipendente su un solo modello ha trovato Qwen3.8-27B in Q4_K_M pari a BF16 su GPQA Diamond, IFBench e Terminal-Bench 2.1 [method-5]; Artificial Analysis ha misurato lo stesso punteggio di indice per i pesi BF16 e NVFP4 di Nemotron 3.5 Lightning [benchmarks-10]; NVIDIA riporta un recupero mediano di accuratezza del 99,62% per i pesi NVFP4 di Nemotron 3.5 Super (dato del produttore) [benchmarks-54] e un degrado dell'1% o meno da FP8 a NVFP4 su DeepSeek-R1-0528 (dato del produttore) [method-2].
- Lettura: dove entrano, i pesi a 8 bit sono la scelta predefinita più sicura; i pesi a 4 bit devono superare la qualifica della sezione 7.5 su prompt in italiano e lunghi. Il calcolatore dà la velocità a entrambe le precisioni, non la qualità. Su un sistema compatto le taglie della sezione 6.1 richiedono pesi a 4 bit.

8.4 Licenze

La questione del licenziatario su un nodo presso il cliente (SparkTech o il cliente) decide se si applicano i tetti di ricavi (Mistral Modified MIT: USD 20 milioni di ricavi mensili), la clausola "AI Work Assistant" della Qwen Community License e la clausola sui dispositivi hardware di Step 5; serve un avvocato [models-59][models-22][models-153]. Per alcuni modelli la classificazione delle licenze differisce tra i tag di Hugging Face, Arena ed Epoch AI (Gemma 4 31B: Apache-2.0 nella scheda, "restricted use" per Epoch AI) [benchmarks-56][benchmarks-14]; il riferimento è il file di licenza di ciascun repository.

8.5 Prezzi e disponibilità

Tra il 2025 e novembre 2026 i prezzi sono saliti da 1,6 a 3,3 volte per diversi articoli (listino USA di DGX Spark +74%, listino di Framework Desktop 128 GB +73%, kit di sviluppo Jetson AGX Thor +57%, prezzo di strada della RTX 5090 3,3 volte il listino UE) [hardware-3][hardware-4][hardware-21][hardware-26][hardware-44][hardware-45]. Lo stesso chip si vende a prezzi molto diversi: il 16 novembre 2026 in Germania un sistema GB10 di ASUS con 1 TB di archiviazione partiva da EUR 5.999 e quello di NVIDIA, con 4

TB, da EUR 11.383,97, mentre il sistema ASUS da 64 GB (2 offerte) costava più di quello da 128 GB [hardware-9][hardware-8][hardware-10]. I prezzi di acquisto degli acceleratori AMD Instinct non sono stati trovati. Ogni offerta richiede un preventivo datato.

8.6 Limiti delle stime e delle misure

- **La taratura è nel campione e scarsa.** I fattori di efficienza e il calcolo efficace vengono da 64 punti di decode a richiesta singola, 56 punti di prefill e 13 serie di concorrenza, molti da singoli operatori della comunità; solo MLPerf (misurato da chi sottopone il risultato e sottoposto a revisione tra pari; le voci DGX Spark e Ryzen AI Max+ 395 sono state sottoposte dal produttore del motore di inferenza usato, la voce Jetson AGX Thor da NVIDIA) e Artificial Analysis (indipendente) sono prove controllate [measurements-26][measurements-24].
- **Dispersione.** L'efficienza di decode misurata va da 0,38 a 0,91 sui sistemi compatti e da 0,21 a 0,82 sulle schede; il calcolo efficace di prefill su GB10 va da 11,9 a 32,4 TFLOPS tra modelli e build (`out/tabelle.md`, tabella 1). Le stime per singola coppia possono sbagliare di decine di punti percentuali; lo scarto per Qwen3.6-35B-A3B su Ryzen AI Max+ 395 è dal 21% al 46% (sezione 5.3).
- **gpt-oss pesa sulla taratura della classe 1.** 16 dei 34 punti di decode dei MoE su memoria unificata sono prove di gpt-oss (11 su 14 su GB10); la loro efficienza mediana è 0,59 contro 0,52 degli altri modelli, e non è stata trovata alcuna misura di Qwen3.6-35B-A3B su GB10. La tabella 12 di `out/tabelle.md` dà le taglie della classe 1 a 0,52, 0,46 e 0,38: il verdetto che si sposta è la taglia M entro gli obiettivi su GB10.
- **Che cosa il modello tralascia:** la decodifica speculativa (che ha portato il decode da 30,3 a 54,9 tok/s per Gemma 4 26B-A4B su DGX Spark [measurements-12]), la cache dei prefissi (che accorcerebbe la documentazione ripetuta di una sessione di codifica), lo stato fisso degli strati ad attenzione lineare, l'effetto della profondità sul prefill, la banda usata dal browser sui sistemi a memoria unificata, le differenze tra tokenizer e i motori diversi da quelli misurati. Le stime di prefill per le schede senza misure (RTX PRO 5000, RTX PRO 4000, W7900, Arc Pro B60, L40S, H200 NVL) sono scalate dal calcolo dichiarato; la RTX PRO 4500 non ha un valore dichiarato da cui scalare e non è stimata.
- **Dati di carico:** 3,5 caratteri per token per l'italiano è un'ipotesi; le quote notturne di rotture, esecuzioni instabili, difetti e modifiche sono ipotesi (solo il 3% usato per la pianificazione e l'1,5% di Google hanno una fonte) [workload-serving-9][workload-serving-63]; i conteggi reali di token sono stati misurati solo su due modelli (sezione 5.7: da 3,4 a 4,3 caratteri per token per i prompt di codifica, riparazione e diagnosi, da 2,6 a 2,8 per il prompt delle statistiche API, e analisi API da 700 a 881 token di risposta contro i 418 ipotizzati nella sezione 1.2) [workload-serving-10].
- **Le raffiche sono una stima per eccesso:** il calcolatore assume che tutte le chiamate concorrenti arrivino insieme, e il caso con la documentazione al limite assume che ogni giro di codifica concorrente porti tutti i 24.000 caratteri.
- **Pagine vive** (InferenceX, Artificial Analysis) possono cambiare dopo il 16 novembre 2026 [measurements-30][measurements-25].

Per il vostro reparto IT: considerate provvisorie le taglie che nella sezione 6 sono vicine a un limite (GB10 alla taglia M entro gli obiettivi, Ryzen AI Max+ 395 a qualunque taglia) finché non sono misurate sul vostro hardware (sezione 7.5).

8.7 Italiano

Non è stata trovata alcuna valutazione specifica in italiano dei modelli open-weight del 2026. Gli aggregati multilingue mostrano uno scarto piccolo (indice italiano di Artificial Analysis: Qwen3.5-397B-A17B 93 contro 95 del primo modello) [benchmarks-13]. L'unico studio del 2026 su ITALIC (10.000 domande da concorsi pubblici italiani) copre modelli fino a 20B: Ministral-3-8B dall'80,2% all'87,4%, Gemma 3 12B dal 76,5% al 79,0%, Qwen3-8B dal 70,8% al 77,4%, gpt-oss-20b dal 26,1% al 76,9% secondo il formato del prompt, contro l'86,1-86,8% di GPT-5 nano [benchmarks-57]. L'italiano compare nelle schede dei modelli Mistral, Nemotron, Granite ed EuroLLM; Qwen, Gemma e Muse dichiarano un'ampia copertura senza nominarlo (tabella 3.1). L'insieme di qualifica della sezione 7.5 deve quindi comprendere test in italiano.

9. Metodo e riproducibilità

9.1 Fonti ed etichette

Gli input sono cinque note di ricerca, compilate il 16 novembre 2026 da fonti primarie dove possibile (schede e file di configurazione dei modelli, specifiche dei produttori, file grezzi dei benchmark, codice sorgente del software di serving): `fonti/models.*` (120 modelli), `fonti/hardware.*` (57 record di prodotti, sistemi e

prezzi), `fonti/measurements.*` (185 righe di misura), `fonti/benchmarks.*` (607 risultati di benchmark, 610 dopo la verifica) e `fonti/workload-serving.*` (carico di SparkTest ricavato dal codice, software di serving). Per le stime di questo studio non sono stati scaricati pesi di modelli e non è stato eseguito alcun modello; la misura della sezione 5.7 è stata fatta dopo, il 17 novembre 2026, ed è indicata come misurata da SparkTech. L'autore ha aggiunto cinque riferimenti per il metodo (da [method-1] a [method-5]). Ogni nota termina con le correzioni fatte dopo la verifica del 17 novembre 2026.

9.2 Formule

GRANDEZZA	FORMULA	ETICHETTA
Pesi	parametri in memoria x bit per peso / 8	stima
Cache KV, attenzione a query raggruppate	2 (K e V) x strati con cache x teste KV x dimensione della testa x byte per valore x token	stima (valori di configurazione dichiarati)
Cache KV, multi-head latent attention	strati con cache x (dimensioni latenti + RoPE) x byte per valore x token	stima
Strati a finestra scorrevole	come sopra, con i token limitati alla finestra	stima
Ingombro	pesi + sessioni x KV(32k) + riserva <= memoria	stima
Limite di decode	banda / (parametri attivi x bit / 8 + KV letta alla profondità corrente)	stima
Stima di decode	limite di decode x efficienza (sezione 9.3)	stima
Stima di prefill	calcolo efficace / (2 x parametri attivi)	stima
Tempo di una chiamata	token di prompt / prefill + token di risposta / decode, con il decode alla profondità prompt + metà della risposta	stima
Raffica di c chiamate	c x token di prompt / prefill + token di risposta / (decode x c^(beta - 1))	stima
Capacità notturna	finestra notturna / (token di prompt per test / prefill + token di risposta per test / (decode x c^beta)), con il decode a una profondità di 5.000 token	stima
Verdetto di taglia	la taglia più grande le cui raffiche di codifica, riparazione e diagnosi rispettano tutte gli obiettivi (o i limiti rigidi) e le cui sessioni concorrenti da 32k entrano in memoria; calcolato con i prompt tipici e con il prompt di codifica con la documentazione al limite	stima
Utilizzabile	il minore dei due verdetti sui limiti rigidi	stima

9.3 Valori di taratura (calcolati dal calcolatore da `dati/measurements.json`)

FATTORE	VALORE	BASE
Efficienza di decode, memoria unificata, denso	0,84 (da 0,67 a 0,91; 11 righe)	mediana di misurato / limite, senza decodifica speculativa, profondità fino a 4.096
Efficienza di decode, memoria unificata, MoE	0,56 (da 0,38 a 0,78; 34 righe); 0,52 senza le 16 righe di gpt-oss (solo per la sensibilità)	idem
Efficienza di decode, scheda GPU, denso	0,67 (da 0,37 a 0,82; 7 righe)	idem
Efficienza di decode, scheda GPU, MoE	0,28 (da 0,21 a 0,39; 12 righe)	idem
Calcolo efficace di prefill, GB10	19,7 TFLOPS MoE (11 righe); 38,9 denso (2 righe)	mediana del prefill misurato x 2 x parametri attivi
Calcolo efficace di prefill, Ryzen AI Max+ 395	6,7 TFLOPS MoE (21 righe); 20,2 denso (7 righe)	idem; il PRO 495 usa questi valori x 61/60 (rapporto dichiarato di FP16 della GPU)
Calcolo efficace di prefill, schede	RTX PRO 6000 69,6 MoE; RTX 5090 56,5 MoE; R9700 22,8 MoE, 58,4 denso; Arc Pro B70 69,7 denso	idem
Rapporto denso / MoE	2,57	piattaforme misurate con entrambi
kappa (calcolo efficace / calcolo denso a 8 bit dichiarato)	0,173 denso, 0,067 MoE (5 schede)	usato per le schede senza misure di prefill
Esponente di concorrenza beta	0,44 MoE (9 serie); 0,96 denso (4 serie)	serie di concorrenza misurate

Una prova della comunità su un sistema Ryzen AI Max+ PRO 495 dà 16,1 TFLOPS di calcolo efficace di prefill per GLM-5.3-Flash, 2,4 volte la mediana di Ryzen AI Max+ 395 [measurements-21]; con una sola riga e un'architettura insolita non viene usata. I due backend di Ryzen AI Max+ 395 aggiunti dopo la verifica (riga T1-41) sono tenuti per confronto e non usati, per non aggiungere altri punti di gpt-oss alla taratura.

9.4 Ipotesi

IPOTESI	VALORE	BASE
Bit per peso	BF16 16; 8 bit 8,5; 4 bit 4,5	llama.cpp Q8_0 8,5008 [method-1]; NVFP4 4,5 [method-2]; i file GGUF pubblici a 4 bit misurano da 4,3 a 5,4 bit per peso (dimensione del file / parametri, [measurements-20])
Precisione della cache KV	BF16	FP8 dimezza i valori KV
Lunghezza della sessione per l'ingombro	32.768 token	contesto minimo raccomandato dallo studio sul carico [workload-serving-10]
Riserva	16 GB su memoria unificata; il 10% della memoria su una GPU	ipotesi di questo studio; come ordine di grandezza, vLLM usa per impostazione predefinita il 92% della memoria GPU quando la sua cache KV non è estensibile [method-3]
Concorrenza	3, 5, 9 richieste per S, M, L	[workload-serving-4][workload-serving-9]
Notte	6 ore; carico di una notte di rilascio, limite alto per test (3.394 token di prompt e 233 di risposta, con la regola del 17 novembre 2026; prima 4.975 e 256)	[workload-serving-10][workload-serving-67]
Unità di memoria	pesi e KV in GB da 10^9 byte; capacità come dichiarate	dove i GB dichiarati sono binari, i byte utilizzabili sono fino al 7% in più

9.5 File e come rieseguire

```
marketing/research/
  fonti/              note di ricerca (Markdown) e loro dati (JSON), 16 novembre 2026, con le
correzioni del 17 novembre 2026
  dati/normalizza.mjs  costruisce dati/*.json da fonti/ (le scelte curate sono scritte nello
script)
  dati/models.json     120 modelli, normalizzati, con le fonti
  dati/hardware.json   39 piattaforme, 18 sistemi, 22 configurazioni confrontate, ipotesi,
fonti
  dati/measurements.json  185 righe di misura associate a modelli e hardware, valori di
quantizzazione, fonti
  dati/workload.json   taglie di SparkTest, operazioni, carichi notturni, dati di serving,
fonti
  dati/sources.json    tutte le 455 fonti
  calcolatore.mjs      il calcolatore (Node, senza dipendenze)
  out/tabelle.md       output del calcolatore, 12 tabelle (in inglese)
  out/tabelle.json     lo stesso, leggibile da programma
  qualifica/           la misura della sezione 5.7: rapporti del kit, script usati sul server
di prova, dati
```

Comandi, da `marketing/research/`: `node dati/normalizza.mjs` e poi `node calcolatore.mjs`. Entrambi girano in molto meno di un secondo, senza accesso alla rete. Ogni numero delle sezioni 5 e 6 si può ricondurre a `out/tabelle.json`; ogni input a un identificativo di fonte in `dati/*.json`.

Riferimenti

Tutte le fonti sono state consultate il 16 novembre 2026, salvo diversa indicazione nella voce. Le voci indicate come interne si riferiscono al repository di SparkTech e non hanno un URL pubblico. Le voci con "ricontrollato il 2026-11-17" sono state riaperte durante la verifica. I titoli delle fonti sono nella lingua originale.

Carico di SparkTest e software di serving (nota `fonti/workload-serving.md`)

- [workload-serving-1] Sorgente dell'agente SparkTest, `agent/src/llm.js` (generatori di prompt, client HTTP, timeout, schema delle analisi API). SparkTech, repository interno, copia di lavoro del 2026-11-16 22:09. Consultato il 2026-11-16. Interno, senza URL pubblico.

- [workload-serving-2] Sorgente dell'agente SparkTest, `agent/src/plan.js` (DSL del piano, `PLAN_SCHEMA`, validazione). SparkTech, interno, copia di lavoro del 2026-11-16 22:02. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-3] Sorgente dell'agente SparkTest, `agent/src/harness.js` (istantanea della pagina, limiti di cattura del traffico di rete). SparkTech, interno, copia di lavoro del 2026-11-16 22:29. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-4] Sorgente dell'agente SparkTest, `agent/src/codify.js`, `agent/src/runner.js`, `agent/src/config.js` (budget di codifica, tentativi di riparazione, diagnosi, ricostruzione dopo un fallimento fino al 2026-11-16, slot paralleli). SparkTech, interno, copia di lavoro del 2026-11-16 22:31. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-5] Sorgente della console SparkTest, `web/lib/suggest-tests.ts`, `web/lib/chat-prompt.ts`, `web/lib/prompt-data.ts`, `web/app/api/portal/chat/route.ts`. SparkTech, interno, copia di lavoro del 2026-11-16 22:40. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-6] Sorgente della console SparkTest, `web/lib/schedule.ts` (suite notturna), `web/lib/api-insights.ts` (analisi su richiesta e relativi limiti), `web/app/api/agent/claim/route.ts` (documentazione inviata all'agente). SparkTech, interno. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-9] Note di prodotto di SparkTest, `docs/VISION.md` (taglie on premise S, M, L: 100, 400, 1.500 test; 2, 4, 8 esecuzioni parallele) e `docs/BUSINESS-MODEL.md` (quota di riparazione del 3% usata per la pianificazione). SparkTech, interno. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-10] Script di misura e output grezzi di questo studio: `marketing/research/dati/workload-serving-prompt-sizes.mjs` (+ `.out.txt`) e `marketing/research/dati/workload-serving-night-load.mjs` (+ `.out.txt`). 2026-11-16. Consultato il 2026-11-16. Interno, senza URL pubblico.
- [workload-serving-11] "What are tokens and how to count them?", OpenAI Help Center, https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them. Lettura diretta rifiutata (HTTP 403); regola empirica letta nella resa della pagina offerta dal motore di ricerca. Consultato il 2026-11-16.
- [workload-serving-12] Note di rilascio di vLLM v0.33.0, progetto vLLM, https://github.com/vllm-project/vllm/releases/tag/v0.33.0 (note pubblicate con v0.33.0rc3 il 2026-11-02). Consultato il 2026-11-16.
- [workload-serving-16] Codice sorgente di vLLM al tag v0.33.0: `vllm/entrypoints/generate/base/protocol.py` (`JsonSchemaResponseFormat`), `vllm/entrypoints/serve/engine/protocol.py` (`OpenAIBaseModel`, campi aggiuntivi), `vllm/entrypoints/openai/chat_completion/protocol.py` (`reasoning_effort`), https://github.com/vllm-project/vllm/tree/v0.33.0/vllm/entrypoints. Consultato il 2026-11-16.
- [workload-serving-17] Codice sorgente di vLLM al tag v0.33.0: `vllm/config/structured_outputs.py` e `vllm/v1/structured_output/backend_xgrammar.py`, https://github.com/vllm-project/vllm/tree/v0.33.0/vllm. Consultato il 2026-11-16.
- [workload-serving-19] "vllm serve" riferimento della CLI, documentazione di vLLM, https://docs.vllm.ai/en/latest/cli/serve.html. Consultato il 2026-11-16.
- [workload-serving-20] "Usage Stats Collection", documentazione di vLLM, https://docs.vllm.ai/en/latest/usage/usage_stats.html. Consultato il 2026-11-16.
- [workload-serving-22] "CUDA Toolkit Release Notes" (versioni del toolkit e versione minima del driver), NVIDIA, https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html. Consultato il 2026-11-16.
- [workload-serving-24] Codice sorgente di SGLang al tag v0.5.23: `python/sglang/srt/entrypoints/openai/protocol.py`, https://github.com/sgl-project/sglang/blob/v0.5.23/python/sglang/srt/entrypoints/openai/protocol.py. Consultato il 2026-11-16.
- [workload-serving-26] "Server Arguments", documentazione di SGLang, https://docs.sglang.io/docs/advanced_features/server_arguments. Consultato il 2026-11-16.
- [workload-serving-29] "llama-server" README, ggml-org/llama.cpp, https://github.com/ggml-org/llama.cpp/blob/master/tools/server/README.md. Consultato il 2026-11-16.
- [workload-serving-30] "GBNF Guide", sezione "JSON Schemas to GBNF", ggml-org/llama.cpp, https://github.com/ggml-org/llama.cpp/blob/master/grammars/README.md. Consultato il 2026-11-16.
- [workload-serving-31] Sorgente di llama.cpp a b12445: `tools/server/server-common.cpp`, https://github.com/ggml-org/llama.cpp/blob/b12445/tools/server/server-common.cpp. Consultato il 2026-11-16.

- [workload-serving-33] "OpenAI compatibility", documentazione di Ollama a v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/api/openai-compatibility.mdx. Consultato il 2026-11-16.
- [workload-serving-35] "Context length", documentazione di Ollama a v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/context-length.mdx. Consultato il 2026-11-16.
- [workload-serving-36] "FAQ" (richieste parallele, coda), documentazione di Ollama a v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/faq.mdx. Consultato il 2026-11-16.
- [workload-serving-37] "Authentication", documentazione di Ollama a v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/api/authentication.mdx. Consultato il 2026-11-16.
- [workload-serving-40] Sorgente di Ollama a v0.40.16: `openai/openai.go`, https://github.com/ollama/ollama/blob/v0.40.16/openai/openai.go. Consultato il 2026-11-16.
- [workload-serving-41] "Structured Output" (compatibilità OpenAI), documentazione di LM Studio, https://lmstudio.ai/docs/developer/openai-compat/structured-output. Consultato il 2026-11-16.
- [workload-serving-43] "LM Studio is free for use at work", blog di LM Studio, https://lmstudio.ai/blog/free-for-work, 2025-07-08. Consultato il 2026-11-16.
- [workload-serving-45] "Offline Operation", documentazione di LM Studio, https://lmstudio.ai/docs/app/offline. Consultato il 2026-11-16.
- [workload-serving-46] Repository di LM Studio `lmstudio-ai/lms` e `lmstudio-ai/lmstudio-js` (licenza MIT); nessun repository pubblico `lmstudio-ai/llmster` (API di GitHub: non trovato), https://github.com/lmstudio-ai. Consultato il 2026-11-16.
- [workload-serving-48] README di text-generation-inference (avviso di modalità di manutenzione), stato del repository (archiviato) e rilasci (v3.3.7, 2025-12-19), Hugging Face, https://github.com/huggingface/text-generation-inference. Consultato il 2026-11-16.
- [workload-serving-49] documentazione "Guidance" di TGI e sorgente del router `router/src/lib.rs` (`GrammarType`, `JsonSchemaConfig`), Hugging Face, https://github.com/huggingface/text-generation-inference/blob/main/router/src/lib.rs. Consultato il 2026-11-16.
- [workload-serving-50] "Release Notes", NVIDIA NIM for LLM and VLM (rilascio 2.1.6), https://docs.nvidia.com/nim/large-language-models/latest/about-nim-llm/release-notes.html, pagina aggiornata il 2026-11-16. Consultato il 2026-11-16.
- [workload-serving-52] "Air-Gap Deployment", NVIDIA NIM for LLM and VLM, https://docs.nvidia.com/nim/large-language-models/latest/deployment/air-gap-deployment.html. Consultato il 2026-11-16.
- [workload-serving-54] "Pricing", NVIDIA AI Enterprise Licensing Guide, https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/pricing.html, aggiornata il 2026-09-02. Consultato il 2026-11-16.
- [workload-serving-57] "OpenAI API chat/completions endpoint", documentazione di OpenVINO Model Server 2026, https://docs.openvino.ai/2026/model-server/ovms_docs_rest_api_chat.html; rilascio 2026.4.1 (2026-10-06) e licenza (Apache-2.0), https://github.com/openvinotoolkit/model_server. Consultato il 2026-11-16.
- [workload-serving-58] S. Geng, H. Cooper, M. Moskal, S. Jenkins, J. Berman, N. Ranchin, R. West, E. Horvitz, H. Nori, "JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models", arXiv:2501.10868v3, https://arxiv.org/abs/2501.10868, 2025-02-27 (v1 2025-01-18). Consultato il 2026-11-16.
- [workload-serving-61] "Environment variables" (`HF_HUB_OFFLINE`, `HF_HUB_DISABLE_TELEMETRY`, `DO_NOT_TRACK`), documentazione dell'Hub di Hugging Face, https://huggingface.co/docs/huggingface_hub/package_reference/environment_variables. Consultato il 2026-11-16.
- [workload-serving-63] J. Micco, "Flaky Tests at Google and How We Mitigate Them", Google Testing Blog, https://testing.googleblog.com/2016/05/flaky-tests-at-google-and-how-we.html, 2016-05-27. Consultato il 2026-11-16.
- [workload-serving-65] "Command-line API" (`NODE_EXTRA_CA_CERTS`, `NODE_USE_SYSTEM_CA`), documentazione di Node.js, https://nodejs.org/api/cli.html. Consultato il 2026-11-16.
- [workload-serving-67] Sorgente dell'agente SparkTest dopo le modifiche del 2026-11-17: `agent/src/runner.js` (nessuna nuova codifica dopo un'esecuzione fallita; agente 0.2.1, commit del 2026-11-17 00:58), `agent/src/llm.js` (controllo dello scenario su valori attesi e dati digitati; richiesta con `json_schema.name`, `chat_template_kwargs` e `reasoning_effort`, commit del 2026-11-17 01:40). SparkTech, interno. Consultato il 2026-11-17. Interno, senza URL pubblico.

Modelli (nota `fonti/models.md`)

- [models-2] Berkeley Function Calling Leaderboard (BFCL) V4, pagina e data_overall.csv. progetto Gorilla, UC Berkeley. https://gorilla.cs.berkeley.edu/leaderboard.html. la pagina indica Last Updated 2026-04-12; CSV: https://gorilla.cs.berkeley.edu/data_overall.csv. Consultato il 2026-11-16.
- [models-3] Alibaba launches Qwen3.5 with open-weight 397B model and broader language support (secondaria). AlternativeTo. https://alternativeto.net/news/2026/2/alibaba-launches-qwen3-5-with-open-weight-397b-model-and-broader-language-support. February 2026. Consultato il 2026-11-16.
- [models-6] Qwen/Qwen3-Coder-Next: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3-Coder-Next. Repository Hugging Face creato il 2026-01-30, ultima modifica 2026-02-03. Consultato il 2026-11-16.
- [models-10] Qwen/Qwen3.5-122B-A10B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.5-122B-A10B. Repository Hugging Face creato il 2026-02-24, ultima modifica 2026-04-24. Consultato il 2026-11-16.
- [models-12] Qwen/Qwen3.6-35B-A3B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Repository Hugging Face creato il 2026-04-15, ultima modifica 2026-11-07. Consultato il 2026-11-16.
- [models-14] Qwen3.6 35B A3B: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-6-35b-a3b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-15] Qwen/Qwen3.8-27B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.8-27B. Repository Hugging Face creato il 2026-08-05, ultima modifica 2026-11-07. Consultato il 2026-11-16.
- [models-16] Qwen3.8 27B: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-8-27b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-17] Open weights models comparison (Intelligence Index v5.0.1). Artificial Analysis. https://artificialanalysis.ai/models/open-source. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-20] Qwen/Qwen3.5-397B-A17B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Repository Hugging Face creato il 2026-02-16, ultima modifica 2026-04-24. Consultato il 2026-11-16.
- [models-21] Qwen/Qwen3.8-Flash-Next: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.8-Flash-Next. Repository Hugging Face creato il 2026-08-24, ultima modifica 2026-11-07. Consultato il 2026-11-16.
- [models-22] Qwen Community License 1.0 (LICENSE di Qwen3.8-Flash-Next). Qwen. https://huggingface.co/Qwen/Qwen3.8-Flash-Next/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-24] Qwen3.8-Flash-Next: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-8-flash-next. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-30] DeepSeek V4 Preview Release. DeepSeek API Docs, News. https://api-docs.deepseek.com/news/news260424/. 2026-04-24. Consultato il 2026-11-16.
- [models-32] deepseek-ai/DeepSeek-V4-Flash-0731: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731. Repository Hugging Face creato il 2026-07-31, ultima modifica 2026-08-01. Consultato il 2026-11-16.
- [models-34] deepseek-ai/DeepSeek-V4.1-Flash: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Repository Hugging Face creato il 2026-09-10, ultima modifica 2026-10-16. Consultato il 2026-11-16.
- [models-35] deepseek-ai/DeepSeek-V4.1-Flash-1017: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017. Repository Hugging Face creato il 2026-10-17, ultima modifica 2026-11-10. Consultato il 2026-11-16.
- [models-36] openai/gpt-oss-20b: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. openai (Hugging Face). https://huggingface.co/openai/gpt-oss-20b. Repository Hugging Face creato il 2025-08-04, ultima modifica 2025-08-26. Consultato il 2026-11-16.
- [models-37] Introducing gpt-oss. OpenAI. https://openai.com/index/introducing-gpt-oss/. 2025-08-05 (data secondo i risultati di ricerca; pagina non aperta). Consultato il 2026-11-16.

- [models-38] gpt-oss-20b: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-20b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-39] openai/gpt-oss-120b: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. openai (Hugging Face). https://huggingface.co/openai/gpt-oss-120b. Repository Hugging Face creato il 2025-08-04, ultima modifica 2025-08-26. Consultato il 2026-11-16.
- [models-40] gpt-oss-120b (High): intelligence, performance and price analysis. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-120b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-44] Llama 4 Community License Agreement (LICENSE). Meta Platforms. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-45] Llama 4 Acceptable Use Policy (USE_POLICY.md). Meta, meta-llama/llama-models su GitHub. https://raw.githubusercontent.com/meta-llama/llama-models/main/models/llama4/USE_POLICY.md. uscita di Llama 4, aprile 2025. Consultato il 2026-11-16.
- [models-48] meta-models/Muse-Glimmer-30B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. meta-models (Hugging Face). https://huggingface.co/meta-models/Muse-Glimmer-30B. Repository Hugging Face creato il 2026-08-09, ultima modifica 2026-10-22. Consultato il 2026-11-16.
- [models-51] Muse Model License, version Muse Spark Open (LICENSE.md). Meta Platforms. https://huggingface.co/meta-models/Muse-Spark-Open-816B-A42B/blob/main/LICENSE.md. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-52] mistralai/Mistral-Small-3.2-24B-Instruct-2506: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Repository Hugging Face creato il 2025-06-19, ultima modifica 2025-12-22. Consultato il 2026-11-16.
- [models-55] mistralai/Ministral-3-14B-Instruct-2512: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. mistralai (Hugging Face). https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512. Repository Hugging Face creato il 2025-10-31, ultima modifica 2026-07-15. Consultato il 2026-11-16.
- [models-56] mistralai/Mistral-Small-4-119B-2603: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Small-4-119B-2603. Repository Hugging Face creato il 2026-01-23, ultima modifica 2026-07-15. Consultato il 2026-11-16.
- [models-57] Mistral Small 4: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/mistral-small-4. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-59] Modified MIT License (LICENSE di Mistral-Medium-3.5-128B). Mistral AI. https://huggingface.co/mistralai/Mistral-Medium-3.5-128B/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-61] Modified MIT License (LICENSE di Devstral-2-123B-Instruct-2512). Mistral AI. https://huggingface.co/mistralai/Devstral-2-123B-Instruct-2512/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-63] Mistral AI Non-Production License (MNPL-0.1). Mistral AI. https://mistral.ai/licences/MNPL-0.1.md. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-64] mistralai/Mistral-Large-3-675B-Instruct-2512: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512. Repository Hugging Face creato il 2025-11-28, ultima modifica 2026-07-15. Consultato il 2026-11-16.
- [models-71] Gemma 4: Expanding the Gemmaverse with Apache 2.0. Google Open Source Blog. https://opensource.googleblog.com/2026/03/gemma-4-expanding-the-gemmaverse-with-apache-20.html. 2026-04-02. Consultato il 2026-11-16.
- [models-72] google/gemma-4-31B-it: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. google (Hugging Face). https://huggingface.co/google/gemma-4-31B-it. Repository Hugging Face creato il 2026-03-11, ultima modifica 2026-07-20. Consultato il 2026-11-16.
- [models-73] Gemma 4 31B: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/gemma-4-31b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.

- [models-74] google/gemma-4-26B-A4B-it: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. google (Hugging Face). https://huggingface.co/google/gemma-4-26B-A4B-it. Repository Hugging Face creato il 2026-03-11, ultima modifica 2026-07-20. Consultato il 2026-11-16.
- [models-75] Gemma 4 26B A4B: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/gemma-4-26b-a4b. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-79] zai-org/GLM-4.7-Flash: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-4.7-Flash. Repository Hugging Face creato il 2026-01-19, ultima modifica 2026-01-29. Consultato il 2026-11-16.
- [models-82] zai-org/GLM-5: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5. Repository Hugging Face creato il 2026-02-11, ultima modifica 2026-08-11. Consultato il 2026-11-16; ricontrollato il 2026-11-17 ("GLM-5 scales from 355B parameters (32B active) to 744B parameters (40B active)").
- [models-85] zai-org/GLM-5.3: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5.3. Repository Hugging Face creato il 2026-08-25, ultima modifica 2026-10-14. Consultato il 2026-11-16; ricontrollato il 2026-11-17 (stesso modello di base di GLM-5.2; nessun conteggio dei parametri nella scheda).
- [models-86] GLM-5.3 License (LICENSE). Z.ai. https://huggingface.co/zai-org/GLM-5.3/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-87] zai-org/GLM-5.3-Flash: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5.3-Flash. Repository Hugging Face creato il 2026-08-25, ultima modifica 2026-10-15. Consultato il 2026-11-16.
- [models-88] GLM-5.3-Flash: pagina del modello. Artificial Analysis. https://artificialanalysis.ai/models/glm-5-3-flash. nessuna data dei dati indicata sulla pagina. Consultato il 2026-11-16.
- [models-100] moonshotai/Kimi-K3.1-Pro: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. moonshotai (Hugging Face). https://huggingface.co/moonshotai/Kimi-K3.1-Pro. Repository Hugging Face creato il 2026-11-02, ultima modifica 2026-11-05. Consultato il 2026-11-16.
- [models-101] Kimi K3.1-Pro License (LICENSE). Moonshot AI. https://huggingface.co/moonshotai/Kimi-K3.1-Pro/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-113] ibm-granite/granite-4.2-30b: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. ibm-granite (Hugging Face). https://huggingface.co/ibm-granite/granite-4.2-30b. Repository Hugging Face creato il 2026-08-07, ultima modifica 2026-09-04. Consultato il 2026-11-16.
- [models-115] NVIDIA Nemotron Open Model License. NVIDIA. https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-nemotron-open-model-license/. v. December 15, 2025. Consultato il 2026-11-16.
- [models-118] nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16. Repository Hugging Face creato il 2026-08-01, ultima modifica 2026-08-24. Consultato il 2026-11-16.
- [models-119] OpenMDW License Agreement 1.1 (LICENSE di NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16). NVIDIA. https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-121] nvidia/NVIDIA-Nemotron-3.5-Bolt-74B-A9B-BF16: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Bolt-74B-A9B-BF16. Repository Hugging Face creato il 2026-10-09, ultima modifica 2026-10-23. Consultato il 2026-11-16.
- [models-122] nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-BF16: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-BF16. Repository Hugging Face creato il 2026-10-12, ultima modifica 2026-10-22. Consultato il 2026-11-16.
- [models-123] nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16. Repository Hugging Face creato il 2026-06-03, ultima modifica 2026-08-24. Consultato il 2026-11-16.
- [models-124] poolside/Laguna-XS-2.1: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. poolside (Hugging Face). https://huggingface.co/poolside/Laguna-XS-2.1. Repository Hugging Face creato il 2026-06-20, ultima modifica 2026-08-19. Consultato il 2026-11-16.

- [models-132] utter-project/EuroLLM-22B-Instruct-2512: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. utter-project (Hugging Face). https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512. Repository Hugging Face creato il 2025-12-05, ultima modifica 2026-07-31. Consultato il 2026-11-16.
- [models-133] swiss-ai/Apertus-v1.5-8B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. swiss-ai (Hugging Face). https://huggingface.co/swiss-ai/Apertus-v1.5-8B. Repository Hugging Face creato il 2026-07-24, ultima modifica 2026-11-03. Consultato il 2026-11-16.
- [models-137] AlmawaveLabs/Velvet2-3B: scheda del modello (README.md), config.json e metadati dell'API dell'Hub. AlmawaveLabs (Hugging Face). https://huggingface.co/AlmawaveLabs/Velvet2-3B. Repository Hugging Face creato il 2026-10-15, ultima modifica 2026-11-09. Consultato il 2026-11-16.
- [models-146] MiniMax M2.7 Non-Commercial License (LICENSE). MiniMax. https://huggingface.co/MiniMaxAI/MiniMax-M2.7/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-153] Step 5 Preview license (LICENSE.txt). StepFun. https://huggingface.co/stepfun-ai/Step-5-Preview/blob/main/LICENSE.txt. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-162] ERNIE 5.5 Flash Base Model License Agreement v1.0 (LICENSE). Baidu. https://huggingface.co/baidu/ERNIE-5.5-Flash-Base/blob/main/LICENSE. testo della licenza come pubblicato nel repository. Consultato il 2026-11-16.
- [models-175] Structured Outputs (vLLM documentation, latest). vLLM project. https://docs.vllm.ai/en/latest/features/structured_outputs.html. documentazione per sviluppatori senza data. Consultato il 2026-11-16.
- [models-176] llama.cpp HTTP server README. ggml-org. https://github.com/ggml-org/llama.cpp/blob/master/tools/server/README.md. senza data, ramo master. Consultato il 2026-11-16.
- [models-177] Qwen4-Max-Preview (secondaria). AI TL;DR. https://ai-tldr.dev/releases/qwen-qwen4-max-preview/. 2026-11-11. Consultato il 2026-11-16.
- [models-178] Qwen4: four variants, no confirmed open weights (secondaria). Tech Insider. https://tech-insider.org/qwen4-four-variants-no-confirmed-open-weights-alibaba-2026/. 2026 (Apsara Conference del 2026-09-22 secondo l'articolo). Consultato il 2026-11-16.

Hardware (nota `fonti/hardware.md`)

- [hardware-1] NVIDIA. "NVIDIA DGX Spark" (pagina del prodotto, specifiche tecniche) https://www.nvidia.com/en-us/products/workstations/dgx-spark/. Consultato il 2026-11-16.
- [hardware-2] NVIDIA. "DGX Spark User Guide: Hardware Overview". https://docs.nvidia.com/dgx/dgx-spark/hardware.html. Consultato il 2026-11-16.
- [hardware-3] NVIDIA Developer Forums. "2/23/2026 Price Change Announcement" (DGX Spark Founders Edition) https://forums.developer.nvidia.com/t/2-23-2026-price-change-announcement/361713. Consultato il 2026-11-16.
- [hardware-4] The Register, Tobias Mann. "Nvidia debuts $4,999 DGX Spark with half the RAM and storage, amid memory crunch". https://www.theregister.com/a/5300622. Consultato il 2026-11-16.
- [hardware-5] ServeTheHome, John Lee. "NVIDIA DGX Spark 64GB launched and big 128GB GB10 price increases". https://www.servethehome.com/nvidia-dgx-spark-64gb-launched-and-big-128gb-gb10-price-increases/. Consultato il 2026-11-16.
- [hardware-6] NVIDIA. "NVIDIA DGX Spark Warranty". https://www.nvidia.com/en-us/support/warranty/dgx-spark/. Consultato il 2026-11-16.
- [hardware-8] geizhals.de. "PNY NVIDIA DGX Spark Founders Edition (DGXSPARK-FOUNEDIT-EU)". https://geizhals.de/pny-nvidia-dgx-spark-dgxspark-founedit-eu-a3552134.html. Consultato il 2026-11-16.
- [hardware-9] geizhals.de. "ASUS Ascent (DGX Spark) GX10-GG0003BN, 128 GB, 1 TB". https://geizhals.de/asus-ascent-gx10-gg0003bn-90ms0371-m00030-a3551306.html. Consultato il 2026-11-16.
- [hardware-10] geizhals.de. "ASUS Ascent (DGX Spark)" panoramica delle varianti (compresa GX10-GG0062BN 64 GB) https://geizhals.de/asus-ascent-v219542.html. Consultato il 2026-11-16.
- [hardware-15] NVIDIA Developer Forums. "GB10 really does hit ~1 PFLOP NVFP4 (2:4 sparse), measured, with an open-source tool to reproduce it". https://forums.developer.nvidia.com/t/gb10-really-does-hit-1-pflop-nvfp4-2-4-sparse-measured-with-an-open-source-tool-to-reproduce-it/373618. Consultato il 2026-11-16.

- [hardware-16] NVIDIA Developer Forums. "GB10 Hardware Baseline: First Direct Measurements and Findings". https://forums.developer.nvidia.com/t/gb10-hardware-baseline-first-direct-measurements-and-findings/367851. Consultato il 2026-11-16.
- [hardware-17] NVIDIA Developer Forums. "The DDR bandwidth is significantly lower than the claimed 273GB/s". https://forums.developer.nvidia.com/t/the-ddr-bandwidth-is-significantly-lower-than-the-claimed-273gb-s/363238. Consultato il 2026-11-16.
- [hardware-18] AMD. "AMD Ryzen AI Max+ 395" (specifiche del prodotto) https://www.amd.com/en/products/processors/laptop/ryzen/ai-300-series/amd-ryzen-ai-max-plus-395.html. Consultato il 2026-11-16.
- [hardware-19] AMD. "AMD Ryzen AI Halo Developer Platform with Ryzen AI Max+ PRO 495 processor" (specifiche tecniche) https://www.amd.com/en/products/processors/desktops/ryzen/ryzen-ai-halo/ryzen-ai-max-plus-pro-495.html. Consultato il 2026-11-16.
- [hardware-21] Hardware Busters, crmaris. "Framework Desktop goes to 192GB with Gorgon Halo and the price nearly doubles to $6,799". https://hwbusters.com/news/framework-desktop-goes-to-192gb-with-gorgon-halo-and-the-price-nearly-doubles-to-6799/. Consultato il 2026-11-16.
- [hardware-22] geizhals.de. "GMKtec EVO-X2 (Ryzen AI Max+ 395, 128 GB, 2 TB)". https://geizhals.de/gmktec-evo-x2-a3485299.html. Consultato il 2026-11-16.
- [hardware-25] NVIDIA. "Jetson Thor" (specifiche di Jetson AGX Thor) https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-thor/. Consultato il 2026-11-16.
- [hardware-26] CNX Software, Jean-Luc Aufranc. "NVIDIA increases the price of Jetson modules and devkits by up to 101%". https://www.cnx-software.com/2026/07/22/nvidia-increases-the-price-of-jetson-modules-and-devkits-by-up-to-101/. Consultato il 2026-11-16.
- [hardware-27] ServeTheHome, Ryan Smith. "NVIDIA RTX Spark N1X launch overview". https://www.servethehome.com/nvidia-rtx-spark-n1x-launch-overview-nvidia-arms-windows/. Consultato il 2026-11-16.
- [hardware-28] Tech My Money, Michael John-Anyaehie. "Surface RTX Spark Dev Box pre-order ($5,999)". https://techmymoney.com/2026/10/07/surface-rtx-spark-dev-box-pre-order-5999/. Consultato il 2026-11-16.
- [hardware-30] NVIDIA. "NVIDIA DGX Station" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/products/workstations/dgx-station/. Consultato il 2026-11-16.
- [hardware-31] ServeTheHome, Ryan Smith. "NVIDIA DGX Station Systems Available At Last: GB300 and GB200 Workstations For Your Desktop". https://www.servethehome.com/nvidia-dgx-station-systems-available-at-last-gb300-gb200-workstations-for-your-desktop/. Consultato il 2026-11-16.
- [hardware-32] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Workstation Edition" (pagina del prodotto) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-6000/. Consultato il 2026-11-16.
- [hardware-33] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition" datasheet (5349650, JUN26) https://www.nvidia.com/content/dam/en-zz/Solutions/products/workstations/professional-desktop-gpus/rtx-pro-6000-max-q/workstation-datasheet-blackwell-rtx-pro-6000-max-q-nvidia-us-5349650-web.pdf. Consultato il 2026-11-16.
- [hardware-34] NVIDIA. "NVIDIA RTX PRO Blackwell GPU Architecture" whitepaper v1.1, appendice A, tabella 4. https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/pdf/NVIDIA-RTX-Blackwell-PRO-GPU-Architecture-v1_1.pdf. Consultato il 2026-11-16.
- [hardware-35] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Server Edition" datasheet (3850600, APR25), copia ospitata da Akamai. https://www.akamai.com/site/en/documents/corporate/nvidia-rtx-pro-6000-blackwell-server-edition-datasheet.pdf. Consultato il 2026-11-16.
- [hardware-36] NVIDIA. "NVIDIA RTX PRO 5000 Blackwell" datasheet (5488550), copia ospitata da Open Zeka. https://openzeka.com/wp-content/uploads/2026/08/workstation-datasheet-blackwell-rtx-pro-5000-5488550-nvidia.pdf. Consultato il 2026-11-16.
- [hardware-37] NVIDIA. "NVIDIA RTX PRO 5000 Blackwell" (pagina del prodotto) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-5000/. Consultato il 2026-11-16.
- [hardware-38] NVIDIA. "NVIDIA RTX PRO 4500 Blackwell" (pagina del prodotto) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-4500/. Consultato il 2026-11-16.
- [hardware-39] NVIDIA. "NVIDIA RTX PRO 4000 Blackwell" (pagina del prodotto) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-4000/. Consultato il 2026-11-16.

- [hardware-42] NVIDIA. "NVIDIA RTX Blackwell GPU Architecture" whitepaper v1.1, appendice A, tabella 3 (GeForce RTX 5090) https://images.nvidia.com/aem-dam/Solutions/geforce/blackwell/nvidia-rtx-blackwell-gpu-architecture.pdf. Consultato il 2026-11-16.
- [hardware-43] NVIDIA. "NVIDIA Driver License Agreement" (GeForce), v. February 25, 2025, sezione 2.8. https://www.nvidia.com/en-us/drivers/geforce-license. Consultato il 2026-11-16.
- [hardware-44] Notebookcheck. "Nvidia senkt Preise für GeForce RTX 5090, RTX 5080 und RTX 5070 in Europa". https://www.notebookcheck.com/Nvidia-senkt-Preise-fuer-GeForce-RTX-5090-RTX-5080-und-RTX-5070-in-Europa-so-teuer-sind-die-Grafikkarten-tatsaechlich.982114.0.html. Consultato il 2026-11-16.
- [hardware-45] geizhals.de. "MSI GeForce RTX 5090" panoramica delle varianti. https://geizhals.de/msi-geforce-rtx-5090-v186742.html. Consultato il 2026-11-16.
- [hardware-47] geizhals.de. "NVIDIA RTX PRO 6000 Blackwell Workstation Edition, 96GB GDDR7, bulk (900-5G144-2200-000)". https://geizhals.de/nvidia-rtx-pro-6000-blackwell-workstation-edition-900-5g144-2200-000-a3474510.html. Consultato il 2026-11-16.
- [hardware-48] geizhals.de. "NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (900-5G153-2500-000)". https://geizhals.de/nvidia-rtx-pro-6000-blackwell-max-q-workstation-edition-900-5g153-2500-000-a3488390.html. Consultato il 2026-11-16.
- [hardware-49] geizhals.de. "NVIDIA RTX PRO 5000 Blackwell, 48GB GDDR7, retail (900-5G153-2550-000)". https://geizhals.de/nvidia-rtx-pro-5000-blackwell-900-5g153-2550-000-a3484625.html. Consultato il 2026-11-16.
- [hardware-50] geizhals.de. "NVIDIA RTX PRO 5000 Blackwell, 72GB GDDR7, retail (900-5G153-2570-000)". https://geizhals.de/nvidia-rtx-pro-5000-blackwell-900-5g153-2570-000-a3687804.html. Consultato il 2026-11-16.
- [hardware-51] geizhals.de. "NVIDIA RTX PRO 4500 Blackwell Workstation Edition, 32GB GDDR7, bulk (900-5G147-2250-000)". https://geizhals.de/nvidia-rtx-pro-4500-blackwell-900-5g147-2250-000-a3473415.html. Consultato il 2026-11-16.
- [hardware-52] geizhals.de. "NVIDIA RTX PRO 4000 Blackwell, 24GB GDDR7, retail (900-5G147-2570-000)". https://geizhals.de/nvidia-rtx-pro-4000-blackwell-900-5g147-2570-000-a3488409.html. Consultato il 2026-11-16.
- [hardware-54] AMD. "AMD Radeon AI PRO R9700" (specifiche del prodotto) https://www.amd.com/en/products/graphics/workstations/radeon-ai-pro/ai-9000-series/amd-radeon-ai-pro-r9700.html. Consultato il 2026-11-16.
- [hardware-55] AMD. "AMD Radeon PRO W7900" (specifiche del prodotto) https://www.amd.com/en/products/graphics/workstations/radeon-pro/w7900.html. Consultato il 2026-11-16.
- [hardware-57] geizhals.de. "XFX Radeon AI PRO R9700". https://geizhals.de/xfx-radeon-ai-pro-r9700-rx-97xproaiy-a3650663.html. Consultato il 2026-11-16.
- [hardware-58] geizhals.de. "AMD Radeon PRO W7900". https://geizhals.de/amd-radeon-pro-w7900-a2948314.html. Consultato il 2026-11-16.
- [hardware-59] Intel. "Intel Arc Pro B-Series Graphics Quick Reference Guide" (aggiornamento di marzo 2026) https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2026-03/intel-arc-pro-b-series-graphics-quick-reference-guide-v1-0.pdf. Consultato il 2026-11-16.
- [hardware-60] StorageReview, Dylan Dougherty. "Intel Arc Pro B70 Review". https://www.storagereview.com/review/intel-arc-pro-b70-review. Consultato il 2026-11-16.
- [hardware-61] geizhals.de. "ASRock Arc Pro B70 Creator 32GB". https://geizhals.de/asrock-arc-pro-b70-creator-32gb-a3772447.html. Consultato il 2026-11-16.
- [hardware-62] geizhals.de. "Sparkle Arc Pro B60 Passive". https://geizhals.de/sparkle-arc-pro-b60-passive-1a1-s00442300g-a3652301.html. Consultato il 2026-11-16.
- [hardware-63] NVIDIA. "NVIDIA L4 Tensor Core GPU" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/l4/. Consultato il 2026-11-16.
- [hardware-64] NVIDIA. "NVIDIA L40S" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/l40s/. Consultato il 2026-11-16.
- [hardware-65] NVIDIA. "NVIDIA H100 Tensor Core GPU" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/h100/. Consultato il 2026-11-16.
- [hardware-66] NVIDIA. "NVIDIA H200" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/h200/. Consultato il 2026-11-16.
- [hardware-67] NVIDIA. "NVIDIA DGX B200" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/dgx-b200/. Consultato il 2026-11-16.

- [hardware-68] NVIDIA. "NVIDIA HGX Platform" (specifiche preliminari di HGX B300, HGX B200, HGX Rubin NVL8) https://www.nvidia.com/en-us/data-center/hgx/. Consultato il 2026-11-16.
- [hardware-69] NVIDIA. "NVIDIA DGX B300" (pagina del prodotto, specifiche) https://www.nvidia.com/en-us/data-center/dgx-b300/. Consultato il 2026-11-16.
- [hardware-71] GCN, Hugo Rojas. "NVIDIA Vera Rubin chips begin shipping to cloud providers as AI factory buildout accelerates". https://gcn.com/?p=20421. Consultato il 2026-11-16.
- [hardware-72] AMD. "AMD Instinct MI300X" (specifiche del prodotto) https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html. Consultato il 2026-11-16.
- [hardware-75] AMD. "AMD Instinct MI355X" (specifiche del prodotto) https://www.amd.com/en/products/accelerators/instinct/mi350/mi355x.html. Consultato il 2026-11-16.
- [hardware-76] AMD. "AMD Instinct MI455X" (specifiche del prodotto) https://www.amd.com/en/products/accelerators/instinct/mi400/mi455x.html. Consultato il 2026-11-16.
- [hardware-77] Intel. "Intel Gaudi 3 AI Accelerator White Paper" (V1 Rev. 2), copia ospitata da Presidio. https://www.presidio.com/wp-content/uploads/2025/06/gaudi-3-ai-accelerator-white-paper.pdf. Consultato il 2026-11-16.
- [hardware-79] Comunicato stampa Intel, ripubblicato da Edge AI and Vision Alliance. "Computex: Intel Accelerates AI Everywhere, Redefines Power, Performance and Affordability". https://www.edge-ai-vision.com/2024/06/computex-intel-accelerates-ai-everywhere-redefines-power-performance-and-affordability/. Consultato il 2026-11-16.
- [hardware-82] geizhals.de. "NVIDIA L4 (900-2G193-0000-001)". https://geizhals.de/nvidia-l4-900-2g193-0000-001-a3464751.html. Consultato il 2026-11-16.
- [hardware-83] geizhals.de. "NVIDIA L40S (900-2G133-0080-000)". https://geizhals.de/nvidia-l40s-900-2g133-0080-000-a3127476.html. Consultato il 2026-11-16.
- [hardware-84] geizhals.at. "NVIDIA RTX PRO 6000 Blackwell Server Edition (900-2G153-0000-000)". https://geizhals.at/nvidia-rtx-pro-6000-blackwell-server-edition-900-2g153-0000-000-a3492247.html. Consultato il 2026-11-16.
- [hardware-86] geizhals.de. "NVIDIA H200 NVL (900-21010-0040-000)". https://geizhals.de/nvidia-h200-nvl-900-21010-0040-000-a3414253.html. Consultato il 2026-11-16.
- [hardware-87] American Compute. "GPU server values" (Rack Report October 2026, dati fino ad agosto 2026) https://www.amcompute.com/gpu-server-values. Consultato il 2026-11-16.
- [hardware-88] American Compute. "B300 vs B200" (prezzi a ottobre 2026) https://www.amcompute.com/compare/b300-vs-b200. Consultato il 2026-11-16.
- [hardware-89] American Compute. "GB300 NVL72 vs HGX B300" (prezzi a ottobre 2026) https://www.amcompute.com/compare/gb300-vs-b300. Consultato il 2026-11-16.
- [hardware-91] The Register, Tobias Mann. cronaca del lancio di Intel Gaudi 3 (Vision 2024) https://www.theregister.com/2024/04/09/intel_gaudi_ai_accelerator/. Consultato il 2026-11-16.
- [hardware-92] CG Channel, Jim Thacker. "NVIDIA launches 72GB RTX PRO 5000 Blackwell GPU". https://www.cgchannel.com/2025/12/nvidia-launches-new-72gb-rtx-pro-5000-blackwell-gpu/. Consultato il 2026-11-16.

Misure di inferenza (nota `fonti/measurements.md`)

- [measurements-2] benches/dgx-spark/dgx-spark.md (build 7941, dati rilevati il 2026-02-05). ggml-org / llama.cpp. https://github.com/ggml-org/llama.cpp/blob/master/benches/dgx-spark/dgx-spark.md. Pubblicato: 2026-02-05 (commit 3795cc1). Consultato il 2026-11-16.
- [measurements-3] benches/dgx-spark/dgx-spark.md al commit 15274c0 (build 6989, dati rilevati il 2025-11-02). ggml-org / llama.cpp. https://github.com/ggml-org/llama.cpp/blob/15274c0/benches/dgx-spark/dgx-spark.md. Pubblicato: 2025-11-10. Consultato il 2026-11-16.
- [measurements-5] guide : running gpt-oss with llama.cpp (GitHub Discussion #15396), tabelle di benchmark nel primo post. manutentori di ggml-org / llama.cpp. https://github.com/ggml-org/llama.cpp/discussions/15396. Pubblicato: 2025-08-18 (le tabelle usano le build da 6199 a 6310). Consultato il 2026-11-16.
- [measurements-7] DGX Spark benchmark spreadsheet collegato alla recensione di LMSYS (letta l'esportazione CSV). LMSYS Org. https://docs.google.com/spreadsheets/d/1SF1u0J2vJ-ou-R_Ry1JZQ0iscOZL8UKHpdVFr85tNLU/edit?usp=sharing. Pubblicato: non indicato (collegato il 2025-10-

13). Consultato il 2026-11-16.

- [measurements-9] How NVIDIA DGX Spark's Performance Enables Intensive AI Tasks. NVIDIA Technical Blog. https://developer.nvidia.com/blog/how-nvidia-dgx-sparks-performance-enables-intensive-ai-tasks. Pubblicato: 2025-10-24. Consultato il 2026-11-16.
- [measurements-11] NVIDIA DGX Spark performance. Ollama blog. https://ollama.com/blog/nvidia-spark-performance. Pubblicato: 2025-10-23. Consultato il 2026-11-16.
- [measurements-12] Measured inference benchmarks on a single DGX Spark: same harness across Ollama, llama.cpp and vLLM (notes + data published). NVIDIA Developer Forums, post della comunità. https://forums.developer.nvidia.com/t/measured-inference-benchmarks-on-a-single-dgx-spark-same-harness-across-ollama-llama-cpp-and-vllm-notes-data-published/379766. Pubblicato: 2026-08-10. Consultato il 2026-11-16.
- [measurements-13] Qwen3.8-Flash-Next, GLM-5.3-Flash, DeepSeek-V4-Flash on 2x DGX Spark - prefill, TTFT, SWE-bench Pro. NVIDIA Developer Forums, post della comunità. https://forums.developer.nvidia.com/t/qwen3-8-flash-next-glm-5-3-flash-deepseek-v4-flash-on-2x-dgx-spark-prefill-ttft-swe-bench-pro/382697. Pubblicato: 2026-09-08 (seguiti del 2026-09-08 e del 2026-09-10). Consultato il 2026-11-16.
- [measurements-20] AMD Strix Halo: Legacy Backend Benchmarks (results.json; build di llama.cpp 9187/9193). progetto Strix Halo Toolboxes (comunità). https://kyuz0.github.io/amd-strix-halo-toolboxes/. Pubblicato: 2026-05-18 (dati rilevati il 2026-05-16). Consultato il 2026-11-16; ricontrollato il 2026-11-17 (results.json: cinque backend).
- [measurements-21] Local LLM Benchmarks: Speed results (data/results.json). Local LLM Benchmarks, progetto Strix Halo Toolboxes (comunità). https://local-llm-benchmarks.dev/speed.html. Pubblicato: continuo; prove usate qui datate dal 2026-09-01 al 2026-10-26. Consultato il 2026-11-16.
- [measurements-24] AA-AgentPerf-Local: Benchmarking local AI agents on laptops and workstations. Artificial Analysis. https://artificialanalysis.ai/articles/aa-agentperf-local. Pubblicato: 2026-09-29. Consultato il 2026-11-16; ricontrollato il 2026-11-17.
- [measurements-25] Laptop & Workstation Inference Systems (pagina dei risultati, prezzi a ottobre 2026). Artificial Analysis. https://artificialanalysis.ai/hardware-inference-stack/laptops-workstations. Pubblicato: non indicato (pagina letta il 2026-11-16). Consultato il 2026-11-16; ricontrollato il 2026-11-17 (sistema AMD indicato come Ryzen AI Max+ 395, 128 GB; prove etichettate "(Reasoning)" e "(Xhigh)").
- [measurements-26] MLPerf Inference v6.1 results (summary.csv e result_summary.json di ciascun risultato sottoposto). MLCommons. https://github.com/mlcommons/inference_results_v6.1. Pubblicato: 2026-09-16. Consultato il 2026-11-16; ricontrollato il 2026-11-17 (summary.csv; result_summary.json delle due prove Edge Agentic di Atlas Inference).
- [measurements-28] MLPerf Inference: gpt-oss-120b benchmark description. MLCommons. https://mlcommons.org/2026/03/mlperf-inference-gpt-oss/. Pubblicato: 2026-03-24. Consultato il 2026-11-16.
- [measurements-29] MLPerf Inference v6.1: 5.7x Per-Accelerator Gains, a 512-GPU Run, and Vera Rubin's First Peer-Reviewed Numbers. StorageReview. https://www.storagereview.com/news/mlperf-inference-v6-1-5-7x-per-accelerator-gains-a-512-gpu-run-and-vera-rubins-first-peer-reviewed-numbers. Pubblicato: 2026-09-16. Consultato il 2026-11-16; ricontrollato il 2026-11-17 (descrive il carico Edge Agentic; riporta 20,1 e 19,63 tok/s, nessun valore per turno).
- [measurements-30] InferenceMAX: Open Source Inference Benchmarking. SemiAnalysis. https://newsletter.semianalysis.com/p/inferencemax-open-source-inference. Pubblicato: 2025-10-09. Consultato il 2026-11-16.
- [measurements-32] Running gpt-oss-120b on H100 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-h100. Pubblicato: pagina viva; prove dal 2026-03-27 al 2026-05-17. Consultato il 2026-11-16.
- [measurements-33] Running gpt-oss-120b on H200 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-h200. Pubblicato: pagina viva; prove dal 2025-12-15 al 2026-05-30. Consultato il 2026-11-16.
- [measurements-34] Running gpt-oss-120b on B200 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-b200. Pubblicato: pagina viva; prove dal 2026-02-05 al 2026-05-31. Consultato il 2026-11-16.
- [measurements-35] Running gpt-oss-120b on MI300X (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-mi300x. Pubblicato: pagina viva; prove dal 2026-03-09 al 2026-04-21. Consultato il 2026-11-16.

- [measurements-36] Running gpt-oss-120b on MI355X (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-mi355x. Pubblicato: pagina viva; prove dal 2026-01-14 al 2026-06-05. Consultato il 2026-11-16.
- [measurements-40] Running Qwen3.5 397B on H100 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-h100. Pubblicato: pagina viva; prove dal 2026-07-05 al 2026-10-20. Consultato il 2026-11-16.
- [measurements-41] Running Qwen3.5 397B on H200 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-h200. Pubblicato: pagina viva; prove dal 2026-03-04 al 2026-10-28. Consultato il 2026-11-16.
- [measurements-42] Running Qwen3.5 397B on B200 (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-b200. Pubblicato: pagina viva; prove dal 2026-02-20 al 2026-11-08. Consultato il 2026-11-16.
- [measurements-43] Running Qwen3.5 397B on MI355X (pagina di prova InferenceX). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-mi355x. Pubblicato: pagina viva; prove dal 2026-02-17 al 2026-11-09. Consultato il 2026-11-16.
- [measurements-51] SGLang vs DwarfStar vs vLLM: DeepSeek 4 on RTX Pro 6000. HelixML. https://helix.ml/blog/running-ds4-on-rtx-pro-6000. Pubblicato: 2026-08-14. Consultato il 2026-11-16.

Benchmark (nota `fonti/benchmarks.md`)

- [benchmarks-1] Artificial Analysis Intelligence Index v5 (articolo di lancio). Artificial Analysis. https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5. Consultato il 2026-11-16.
- [benchmarks-3] Open weights models comparison (Intelligence Index v5.0.1, parametri). Artificial Analysis. https://artificialanalysis.ai/models/open-source. Consultato il 2026-11-16.
- [benchmarks-4] Comparison of AI Models; Models leaderboard (Intelligence Index v5.0.1). Artificial Analysis. https://artificialanalysis.ai/models ; https://artificialanalysis.ai/leaderboards/models. Consultato il 2026-11-16.
- [benchmarks-5] Pagine dei modelli: gpt-oss-120b, gpt-oss-20b, Gemma 4 31B, Qwen3.6 35B A3B. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-120b ; https://artificialanalysis.ai/models/gpt-oss-20b ; https://artificialanalysis.ai/models/gemma-4-31b ; https://artificialanalysis.ai/models/qwen3-6-35b-a3b. Consultato il 2026-11-16.
- [benchmarks-10] NVIDIA launches Nemotron 3.5 Lightning. Artificial Analysis. https://artificialanalysis.ai/articles/nemotron-3-5-lightning-launch. Consultato il 2026-11-16.
- [benchmarks-11] Changelog (voci dal 30 ottobre al 16 novembre 2026). Artificial Analysis. https://artificialanalysis.ai/changelog. Consultato il 2026-11-16.
- [benchmarks-13] Multilingual Index: Italian; Multilingual Index overview (basato su Global-MMLU-Lite). Artificial Analysis. https://artificialanalysis.ai/models/multilingual/italian ; https://artificialanalysis.ai/models/multilingual. Consultato il 2026-11-16.
- [benchmarks-14] Capabilities & benchmarking hub, LLM benchmark data (benchmark_data.zip, CC BY 4.0); file usati: swe_bench_verified.csv (eseguito da Epoch), terminalbench_external.csv, os_world_external.csv, osworld_2_external.csv, model_metadata.csv. Epoch AI. https://epoch.ai/benchmarks ; https://epoch.ai/data/benchmark_data.zip. Consultato il 2026-11-16.
- [benchmarks-15] Epoch Capabilities Index (ECI), eci_scores.csv nell'archivio dei dati dei benchmark. Epoch AI. https://epoch.ai/eci ; https://epoch.ai/data/benchmark_data.zip. Consultato il 2026-11-16; ricontrollato il 2026-11-17 (eci_scores.csv datato 2026-11-16).
- [benchmarks-17] Terminal-Bench 4.0 (mini-swe-agent, avg@3). Vals AI. https://www.vals.ai/benchmarks/terminal-bench-4. Consultato il 2026-11-16.
- [benchmarks-18] SWE-bench Verified (archiviato: non più eseguito sui nuovi modelli). Vals AI. https://www.vals.ai/benchmarks/swebench. Consultato il 2026-11-16.
- [benchmarks-23] Classifiche SWE-bench e file di dati leaderboards.json (classifica bash-only tolta dal sito il 2026-09-01). Manutentori di SWE-bench. https://www.swebench.com/ ; https://github.com/SWE-bench/swe-bench.github.io. Consultato il 2026-11-16.
- [benchmarks-24] Why we no longer evaluate SWE-bench Verified (l'accesso automatico ha restituito HTTP 403; contenuto come riassunto in benchmarks-25). OpenAI. https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/. Consultato il 2026-11-16.
- [benchmarks-25] Sintesi del ritiro di SWE-bench Verified (fonte secondaria). Pebblous blog. https://blog.pebblous.ai/blog/swe-bench-verified-retired/en/. Consultato il 2026-11-16.
- [benchmarks-32] Berkeley Function Calling Leaderboard (BFCL) V4 e data_overall.csv. UC Berkeley (progetto Gorilla) https://gorilla.cs.berkeley.edu/leaderboard.html. Consultato il 2026-11-16.

- [benchmarks-34] Risultati ufficiali di OSWorld-Verified (osworld_verified_results.xlsx, valutati dal gruppo di OSWorld con impostazioni uniformi). XLANG Lab, University of Hong Kong. https://osworld-v1.xlang.ai/. Consultato il 2026-11-16.
- [benchmarks-42] Qwen3.8-27B scheda del modello. Qwen team (Alibaba) https://huggingface.co/Qwen/Qwen3.8-27B. Consultato il 2026-11-16.
- [benchmarks-45] Qwen3.6-35B-A3B scheda del modello. Qwen team (Alibaba) https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Consultato il 2026-11-16.
- [benchmarks-48] Muse Glimmer-30B scheda del modello. Meta. https://huggingface.co/meta-models/Muse-Glimmer-30B. Consultato il 2026-11-16.
- [benchmarks-49] gpt-oss-120b & gpt-oss-20b Model Card (arXiv 2508.10925). OpenAI. https://arxiv.org/abs/2508.10925. Consultato il 2026-11-16.
- [benchmarks-54] Nemotron 3.5 Super and Lightning technical report (tabella 10, sezioni 3.8 e 1) e pagina del progetto. NVIDIA. https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3.5-Super-and-Lightning-Technical-Report.pdf ; https://research.nvidia.com/labs/nemotron/Nemotron-3.5-Super-and-Lightning/. Consultato il 2026-11-16.
- [benchmarks-55] Holo3-35B-A3B scheda del modello. H Company. https://huggingface.co/Hcompany/Holo3-35B-A3B. Consultato il 2026-11-16.
- [benchmarks-56] Metadati dei modelli dall'API dell'Hub di Hugging Face (tag di licenza, conteggio dei parametri safetensors, data di creazione del repository) per i repository della sezione 3. Hugging Face. https://huggingface.co/api/models/{repo_id}. Consultato il 2026-11-16.
- [benchmarks-57] Benchmarking EngGPT2-16B-A3B against Comparable Italian and International Open-source LLMs (arXiv 2605.07731v2), tabelle 16 e 17. ENGINEERING Ingegneria Informatica. https://arxiv.org/abs/2605.07731. Consultato il 2026-11-16.

Riferimenti di metodo aggiunti dall'autore

- [method-1] llama.cpp, tools/quantize/README.md: bit per peso di ogni tipo di quantizzazione misurati su Llama-3.1-8B (Q4_K_M 4,8944, Q8_0 8,5008, IQ4_XS 4,4597, F16 16,0005). ggml-org. https://github.com/ggml-org/llama.cpp/blob/master/tools/quantize/README.md. Consultato il 2026-11-16.
- [method-2] E. Alvarez et al., "Introducing NVFP4 for Efficient and Accurate Low-Precision Inference", NVIDIA Technical Blog, 2025-06-24 (NVFP4: una scala FP8 E4M3 ogni 16 valori, 4,5 bit per valore; MXFP4: una scala potenza di due ogni 32 valori). Pubblicazione del produttore. https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/. Consultato il 2026-11-16.
- [method-3] Documentazione di vLLM, riferimento della CLI "vllm serve": --gpu-memory-utilization vale 1,0 per impostazione predefinita con la cache KV estensibile e 0,92 altrimenti. https://docs.vllm.ai/en/latest/cli/serve.html. Consultato il 2026-11-16.
- [method-4] A. Mekala, A. Atmakuru, Y. Song, M. Karpinska, M. Iyyer, "Does quantization affect models' performance on long-context tasks?", arXiv:2505.20276 (v1 2025-05-26, v3 2025-09-20). Studio accademico indipendente. https://arxiv.org/abs/2505.20276. Consultato il 2026-11-16.
- [method-5] P. Migdał, "Benchmarking Qwen3.8 27B quantizations: 4-bit holds up, 1-bit collapses", Quesma blog, 2026-08-26 (llama.cpp, BF16 contro Q8_0 contro Q4_K_M e inferiori, GPQA Diamond, IFBench, Terminal-Bench 2.1). Indipendente, un solo modello, blog. https://quesma.com/blog/qwen38-27b-quantizations-benchmarked/. Consultato il 2026-11-16.
- [method-6] SparkTech, "SparkTest sui modelli open-weight: misure su un server di prova senza GPU", 17 novembre 2026, documento di accompagnamento di questo studio (sezione 5.7). Misurato da SparkTech; dati grezzi e script disponibili su richiesta. https://autotest.sparktech.it/docs/sparktest-open-weight-measurements-it.pdf.

Verifica

Il 16 e il 17 novembre 2026 due revisioni indipendenti hanno controllato questo studio; le correzioni sono state applicate il 17 novembre 2026.

- **Dati e fonti.** 91 affermazioni, che coprono più di 250 valori singoli, sono state confrontate con le fonti primarie citate e, dove possibile, con una seconda fonte; sono stati aperti tutti i 218 URL distinti dei riferimenti (nessuno inesistente; 6 rifiutavano i client automatici e 6 hanno raggiunto un limite di frequenza, e questi sono stati verificati per altra via). Esito: 82 confermate, 7 errate, 1 non verificabile nella fonte citata, 1 con una fonte che non contiene i valori. Tutte e 9 sono state corrette: tre valori

dell'Epoch Capabilities Index riportati come non trovati (tabella 3.2), l'intervallo di SWE-bench Verified della sezione 8.2, l'etichetta dei parametri attivi di GLM-4.7-Flash, il valore FP4 della RTX PRO 5000 da 48 GB, l'intervallo di prefill di Ryzen AI Max+ 395 a 32k di profondità (da 234 a 653 tok/s, cinque backend invece di tre), la fonte dei valori di MLPerf Edge Agentic (tabella 5.6) e la fonte dei conteggi dei parametri di GLM-5.3. Nessuna di queste correzioni cambia una conclusione.

- **Coerenza, etichette e forma.** Il calcolatore è stato rieseguito (output identico), le tabelle 5.1, 5.2, 5.3, 6 e 9.3 sono state confrontate con esso cella per cella (nessuna differenza), cinque righe della tabella 6 sono state ricalcolate con un'implementazione indipendente delle formule e le 228 citazioni sono state confrontate con i riferimenti. Sono stati corretti 34 punti: 17 problemi di coerenza (tra cui le 3 incoerenze interne trovate anche dalla prima revisione), 2 descrizioni di formule mancanti, 6 problemi di etichetta, 8 di unità o di forma e 1 frase sullo SparkNode. Aggiunte per la chiarezza: il riquadro per i reparti IT, il glossario, la scelta per taglia della sezione 6.4 e brevi note nelle sezioni da 1 a 8.
- **Che cosa è cambiato nelle raccomandazioni.** Ogni verdetto di taglia ora controlla anche i prompt di codifica con la documentazione dell'applicazione al limite, e la tabella 6 indica fino a quale taglia è utilizzabile ogni configurazione. Classe 1: Ryzen AI Max+ 395 è utilizzabile fino alla taglia M, non L; su un sistema compatto i modelli da 117B a 125B arrivano alla taglia M su GB10 e alla taglia S su Ryzen AI Max+ 395 (solo gpt-oss-120b), entro i limiti rigidi; con la documentazione al limite, GB10 e le schede da 32 GB rispettano gli obiettivi fino alla taglia S. La tabella 12 di `out/tabelle.md` mostra come si spostano le taglie della classe 1 al variare dell'efficienza di decode.

In totale sono stati corretti 43 punti, a fronte delle 91 affermazioni, delle 5 tabelle, delle 228 citazioni e dei 218 URL controllati.

Aggiornamenti del 17 novembre 2026, dopo la verifica.

- **Nessuna nuova codifica dopo un'esecuzione fallita; le riparazioni tengono i valori.** L'agente non codifica più di nuovo un test dopo un'esecuzione fallita, e il suo controllo dello scenario ora respinge anche una riparazione che cambia un valore atteso o un valore digitato [workload-serving-67]. La sezione 1 è stata allineata; gli scenari notturni sono stati ricalcolati (la notte di rilascio con 1.500 test è passata da 833-1.635 chiamate a 593-1.155) e il calcolatore è stato rieseguito: è cambiata solo la capacità notturna (ultima colonna della tabella 6, sezioni 6.1, 6.2 e 9.4), nessun verdetto sulle taglie.
- **La richiesta a un endpoint del cliente.** La richiesta dell'agente ora dà un nome al suo schema JSON e porta `chat_template_kwargs` e `reasoning_effort`: la sintesi e le sezioni da 7.1 a 7.4 sono state allineate.
- **La misura di SparkTech.** La sezione 5.7 e il punto 7 della sintesi aggiungono la prima misura dei compiti di SparkTest su modelli open-weight, controllata da una verifica indipendente prima della pubblicazione [method-6].