

parametri llama.cpp

Tag: llama.cpp llama llm

Ultimo aggiornamento: 2026-09-17

Guida ai Parametri di llama-server

1. Common Params (Parametri Generali e Risorse)

Informazioni e Utilizzo

Parametro	Descrizione
-h, --help, --usage	Mostra il messaggio di aiuto ed esce dal programma.
--version	Mostra la versione del binario e le informazioni sulla compilazione.
-cl, --cache-list	Elenca i modelli attualmente salvati nella cache locale.
--completion-bash	Genera lo script Bash per abilitare l'autocompletamento dei comandi da terminale.

Gestione Processori e Thread (CPU principali)

Parametro	Descrizione
-t, --threads N	Numero di thread CPU da usare durante la generazione (default: -1).
-tb, --threads-batch N	Numero di thread CPU da usare durante l'elaborazione del batch e del prompt (default: uguale a --threads).
-C, --cpu-mask M	Maschera esadecimale di affinità CPU per limitare l'esecuzione a core specifici.
-Cr, --cpu-range lo-hi	Intervallo di core CPU per l'affinità (es. 0-7).
--cpu-strict <0 1>	Se impostato su 1, forza il posizionamento rigido sui core assegnati (default: 0).
--prio N	Imposta la priorità del processo/thread: -1 (low), 0 (normal), 1 (medium), 2 (high), 3 (realtime).
--poll <0...100>	Grado di polling (attesa attiva) sui thread per ridurre la latenza (default: 50).

Gestione Processori e Thread (Per la fase Batch)

Parametro	Descrizione
-Cb, --cpu-mask-batch M	Maschera CPU specifica per l'elaborazione batch.
-Crb, --cpu-range-batch lo-hi	Intervallo di CPU specifico per l'elaborazione batch.
--cpu-strict-batch <0 1>	Posizionamento CPU rigido per la fase batch.
--prio-batch N	Priorità dei thread durante la fase batch (0-3).
--poll-batch <0 1>	Abilita o disabilita il polling durante la fase batch.

Memoria del Modello e Contesto

Parametro	Descrizione
-c, --ctx-size N	Dimensione della finestra di contesto in token (0 = caricato dal modello).
-n, --predict, --n-predict N	Numero di token da predire (-1 = infinito).
-b, --batch-size N	Dimensione massima logica del batch (default: 2048).
-ub, --ubatch-size N	Dimensione massima fisica del batch / micro-batch (default: 512).
--keep N	Numero di token del prompt iniziale da mantenere in memoria (-1 = tutti).
--swa-full	Utilizza la cache a dimensione intera per i modelli con Sliding Window Attention.
-fa, --flash-attn [on off auto]	Abilita o disabilita Flash Attention (on, off, auto).
--perf, --no-perf	Abilita o disabilita i tempi delle prestazioni interne di libllama.
-e, --escape, --no-escape	Abilita la conversione delle sequenze di escape (\n, \t, ecc.).

Scaling del Contesto (RoPE e YaRN)

Parametro	Descrizione
--rope-scaling {none,linear,yarn}	Metodo di scaling della frequenza RoPE.
--rope-scale N	Estende il contesto di un fattore \$N\$.
--rope-freq-base N	Frequenza base RoPE usata dallo scaling NTK-aware.
--rope-freq-scale N	Fattore di scala della frequenza RoPE (estende il contesto di \$1/N\$).

Parametro	Descrizione
--yarn-orig-ctx N	Dimensione del contesto originale di addestramento del modello per l'algoritmo YaRN.
--yarn-ext-factor N	Fattore di miscela per l'estrapolazione in YaRN.
--yarn-attn-factor N	Scala la magnitudo dell'attenzione o $\sqrt{t}$ in YaRN.
--yarn-beta-slow N	Parametro per la correzione delle dimensioni ad alte frequenze in YaRN.
--yarn-beta-fast N	Parametro per la correzione delle dimensioni a basse frequenze in YaRN.

Gestione Memoria di Sistema e Cache KV

Parametro	Descrizione
-kvo, --kv-offload, -nkvo, --no-kv-offload	Abilita o disabilita il caricamento offload della cache KV.
--repack, -nr, --no-repack	Abilita o disabilita il riimpacchettamento dei pesi in memoria.
--no-host	Bypassa il buffer host consentendo l'uso di buffer extra.
-ctk, --cache-type-k TYPE	Tipo di dati per la chiave (K) nella cache KV (f32, f16, bf16, q8_0, q4_0, ecc.).
-ctv, --cache-type-v TYPE	Tipo di dati per il valore (V) nella cache KV (f32, f16, bf16, q8_0, q4_0, ecc.).
-dt, --defrag-thold N	(DEPRECATO) Soglia di deframmentazione della cache KV.
--mlock	(DEPRECATO) Forza il sistema a mantenere il modello in RAM senza swapping.
--mmap, --no-mmap	(DEPRECATO) Abilita o disabilita la mappatura della memoria per il modello.
-dio, --direct-io, -ndio, --no-direct-io	(DEPRECATO) Utilizza DirectIO se disponibile.
-lm, --load-mode MODE	Modalità di caricamento del modello: auto, none, mmap, mlock, mmap+mlock, dio.
--numa TYPE	Ottimizzazioni per architetture NUMA: distribute, isolate, numactl.

Gestione GPU e Offload Hardware

Parametro	Descrizione
<code>-dev, --device <dev1,dev2,...></code>	Elenco di dispositivi da utilizzare per l'offloading (none per disabilitare).
<code>--list-devices</code>	Stampa l'elenco dei dispositivi disponibili ed esce.
<code>-ot, --override-tensor ...</code>	Sovrascrive il tipo di buffer per uno specifico pattern di tensore.
<code>-cmoe, --cpu-moe</code>	Mantiene tutti i pesi dei modelli Mixture of Experts (MoE) nella CPU.
<code>-ncmoe, --n-cpu-moe N</code>	Mantiene i pesi MoE dei primi \$N\$ layer nella CPU.
<code>-ngl, --gpu-layers N</code>	Numero massimo di layer da memorizzare nella VRAM (auto, all o un numero).
<code>-sm, --split-mode ...</code>	Modalità di suddivisione su più GPU: none, layer, row, tensor.
<code>-ts, --tensor-split N0,N1...</code>	Frazione del modello da allocare su ciascuna GPU (es. 3,1).
<code>-mg, --main-gpu INDEX</code>	GPU principale usata per i calcoli intermedi o con <code>--split-mode row</code> .
<code>-fit [on off]</code>	Regola automaticamente i parametri non definiti per rientrare nella memoria del dispositivo.
<code>-fitt, --fit-target MiB0,MiB1...</code>	Margine di memoria libero target per ciascun dispositivo per l'opzione <code>--fit</code> .
<code>-fitc, --fit-ctx N</code>	Dimensione minima del contesto impostabile tramite l'opzione <code>--fit</code> .

Modifiche, LoRA e Modelli

Parametro	Descrizione
<code>--check-tensors</code>	Controlla i dati dei tensori del modello per individuare valori non validi.
<code>--override-kv KEY=TYPE:VALUE</code>	Sovrascrive i metadati interni del modello per chiave (tipi: int, float, bool, str).
<code>--op-offload, --no-op-offload</code>	Abilita o disabilita l'offload delle operazioni sui tensori host al dispositivo.
<code>--lora FNAME</code>	Percorso dell'adattatore LoRA (valori separati da virgola per caricarne più di uno).
<code>--lora-scaled FNAME:SCALE</code>	Carica adattatori LoRA con un fattore di scala personalizzato.
<code>--control-vector FNAME</code>	Aggiunge uno o più Control Vector al modello.

Parametro	Descrizione
<code>--control-vector-scaled FNAME:SCALE</code>	Aggiunge Control Vector con un fattore di scala personalizzato.
<code>--control-vector-layer-range START END</code>	Intervallo inclusivo di layer su cui applicare i Control Vector.

Download e Sorgente del Modello

Parametro	Descrizione
<code>-m, --model FNAME</code>	Percorso del file locale del modello da caricare.
<code>-mu, --model-url MODEL_URL</code>	URL diretto per il download del modello.
<code>-dr, --docker-repo ...</code>	Repository di modelli Docker Hub (es. gemma3).
<code>-hf, --hf-repo <user>/<model>[:quant]</code>	Repository di Hugging Face (es. ggml-org/GLM-4.7-Flash-GGUF:Q4_K_M).
<code>-hff, --hf-file FILE</code>	Nome del file del modello da scaricare da Hugging Face.
<code>-hft, --hf-token TOKEN</code>	Access Token di Hugging Face per modelli privati o protetti.

Logging

Parametro	Descrizione
<code>--log-disable</code>	Disabilita completamente i log.
<code>--log-file FNAME</code>	Reindirizza l'output dei log su un file specificato.
<code>--log-colors [on off auto]</code>	Abilita o disabilita i log a colori nel terminale.
<code>-v, --verbose, --log-verbose</code>	Imposta il livello di dettaglio dei log al massimo per il debugging.
<code>--offline</code>	Forza la modalità offline prevenendo qualsiasi accesso alla rete.
<code>-lv, --verbosity N</code>	Imposta la soglia di verbosità (0 = generico, 1 = error, 2 = warning, 3 = info, 4 = trace, 5 = debug).
<code>--log-prefix, --no-log-prefix</code>	Abilita o disabilita il prefisso nei messaggi di log.
<code>--log-timestamps, --no-log-timestamps</code>	Abilita o disabilita il timestamp nei messaggi di log.

Cache KV per Modello Draft

Parametro	Descrizione
--spec-draft-type-k, -ctkd	Tipo di dati della chiave (K) nella cache KV per il modello draft.
--spec-draft-type-v, -ctvd	Tipo di dati del valore (V) nella cache KV per il modello draft.

2. Sampling Params (Parametri di Campionamento)

Campionatori Generali

Parametro	Descrizione
--samplers SAMPLERS	Elenco di campionatori da usare ordinati e separati da ;.
-s, --seed SEED	Seme per il generatore di numeri casuali (-1 = seme casuale).
--sampler-seq SEQUENCE	Sequenza abbreviata per definire la catena dei campionatori (default: edskypmxt).
--ignore-eos	Ignora il token di fine sequenza (EOS) e continua la generazione.
--temp, --temperature N	Controlla la casualità dell'output (default: 0.80).
--top-k N	Campionamento Top-K (0 = disabilitato).
--top-p N	Campionamento Top-P / Nucleus (1.0 = disabilitato).
--min-p N	Campionamento Min-P (0.0 = disabilitato).
--top-nsigma N	Campionamento Top-N-Sigma (-1.0 = disabilitato).
--xtc-probability N	Probabilità per l'algoritmo XTC (0.0 = disabilitato).
--xtc-threshold N	Soglia per l'algoritmo XTC (1.0 = disabilitato).
--typical, --typical-p N	Campionamento Locally Typical, parametro p (1.0 = disabilitato).
--repeat-last-n N	Ultimi \$N\$ token da considerare per l'applicazione delle penalità (default: 64).
--repeat-penalty N	Penalità per la sequenza di token ripetuti (1.0 = disabilitato).
--presence-penalty N	Penalità di presenza per la ripetizione dei token (0.0 = disabilitato).
--frequency-penalty N	Penalità di frequenza per la ripetizione dei token (0.0 = disabilitato).

Algoritmo DRY (Don't Repeat Yourself)

Parametro	Descrizione
--dry-multiplier N	Moltiplicatore per il campionamento DRY (0.0 = disabilitato).
--dry-base N	Valore base del campionamento DRY (default: 1.75).
--dry-allowed-length N	Lunghezza massima di sequenza consentita prima che scatti la penalità DRY.
--dry-penalty-last-n N	Numero di token recenti a cui applicare la penalità DRY (0 = disabilita).
--dry-sequence-breaker STRING	Stringa per resettare il controllo del DRY (sostituisce i breaker di default \n, :, ", *).

Campionatori Avanzati

Parametro	Descrizione
--adaptive-target N	Adaptive-P: seleziona i token prossimi a questa probabilità (0.0 a 1.0).
--adaptive-decay N	Adaptive-P: tasso di decadimento per l'adattamento del target nel tempo (0.0 a 0.99).
--dynatemp-range N	Intervallo per la temperatura dinamica (0.0 = disabilitato).
--dynatemp-exp N	Esponente per il calcolo della temperatura dinamica (default: 1.00).
--mirostat N	Attiva il campionamento Mirostat (0 = disattivato, 1 = Mirostat, 2 = Mirostat 2.0).
--mirostat-lr N	Tasso di apprendimento per Mirostat (parametro eta).
--mirostat-ent N	Entropia target per Mirostat (parametro tau).
-l, --logit-bias TOKEN_ID(+/-)BIAS	Modifica manualmente la probabilità di comparsa di uno specifico token.

Vincoli Strutturati e Backend

Parametro	Descrizione
--grammar GRAMMAR	Grammatica in formato tipo BNF per vincolare la generazione del testo.
--grammar-file FNAME	File contenente la grammatica da applicare.
-j, --json-schema SCHEMA	Schema JSON da applicare come vincolo alla generazione.

Parametro	Descrizione
-jf, --json-schema-file FILE	File contenente uno schema JSON da applicare come vincolo.
-bs, --backend-sampling	Delega le operazioni di campionamento al backend hardware (sperimentale).

3. Speculative Params (Decodifica Speculativa)

Configurazione Modello Draft e Hardware

Parametro	Descrizione
--spec-draft-hf ...	Repository Hugging Face da cui scaricare il modello draft.
--spec-draft-threads N	Thread CPU da usare durante la generazione del modello draft.
--spec-draft-threads-batch N	Thread CPU per il batch processing del modello draft.
--spec-draft-cpu-mask M	Maschera CPU di affinità per il modello draft.
--spec-draft-cpu-range lo-hi	Intervallo di CPU per l'affinità del modello draft.
--spec-draft-cpu-strict <0 1>	Posizionamento rigido delle CPU per il modello draft.
--spec-draft-prio N	Priorità del processo/thread del modello draft (0-3).
--spec-draft-poll <0 1>	Polling per l'attesa del lavoro del modello draft.
--spec-draft-cpu-mask-batch M	Maschera CPU batch per il modello draft.
--spec-draft-cpu-strict-batch <0 1>	Posizionamento rigido CPU batch per il modello draft.
--spec-draft-prio-batch N	Priorità processore per la fase batch del draft.
--spec-draft-poll-batch <0 1>	Polling per l'attesa del lavoro batch del draft.
--spec-draft-override-tensor ...	Sovrascrive il buffer di un tensore per il modello draft.
--spec-draft-cpu-moe	Mantiene tutti i pesi MoE del modello draft sulla CPU.
--spec-draft-n-cpu-moe N	Mantiene i pesi MoE dei primi \$N\$ layer del draft sulla CPU.
--spec-draft-device ...	Dispositivi/GPU da usare per l'offloading del modello draft.
--spec-draft-n-gl N	Numero massimo di layer del modello draft da salvare in VRAM.

Parametro	Descrizione
--spec-draft-model FNAME	Percorso del file del modello draft per il campionamento speculativo.

Parametri di Algoritmo Speculativo

Parametro	Descrizione
--spec-draft-n-max N	Numero massimo di token da proporre ad ogni passo (default: 3).
--spec-draft-n-min N	Numero minimo di token draft da utilizzare (default: 0).
--spec-draft-p-split P	Probabilità di split nella decodifica speculativa (default: 0.10).
--spec-draft-p-min P	Probabilità minima per accettare la proposta speculativa (default: 0.00).
--spec-draft-backend-sampling	Attiva o disattiva l'offload del campionamento draft sul backend hardware.
--spec-type ...	Tipi di decodifica speculativa da attivare (es. draft-simple, ngram-simple, ngram-map-k, draft-eagle3).

Parametri N-Gram Speculativi

Parametro	Descrizione
--spec-ngram-mod-n-min N	Numero minimo di token per la decodifica speculativa basata su N-Grammi (default: 48).
--spec-ngram-mod-n-max N	Numero massimo di token per la decodifica speculativa basata su N-Grammi (default: 64).
--spec-ngram-mod-n-match N	Lunghezza di ricerca dell'algoritmo ngram-mod (default: 24).
--spec-ngram-simple-size-n N	Lunghezza dell'N-gramma di ricerca per ngram-simple (default: 12).
--spec-ngram-simple-size-m N	Lunghezza dell'M-gramma bozza per ngram-simple (default: 48).
--spec-ngram-simple-min-hits N	Numero minimo di match richiesti per ngram-simple (default: 1).
--spec-ngram-map-k-size-n N	Lunghezza dell'N-gramma di ricerca per ngram-map-k (default: 12).
--spec-ngram-map-k-size-m N	Lunghezza dell'M-gramma bozza per ngram-map-k (default: 48).

Parametro	Descrizione
--spec-ngram-map-k-min-hits N	Numero minimo di match richiesti per ngram-map-k (default: 1).
--spec-ngram-map-k4v-size-n N	Lunghezza dell'N-gramma di ricerca per ngram-map-k4v (default: 12).
--spec-ngram-map-k4v-size-m N	Lunghezza dell'M-gramma bozza per ngram-map-k4v (default: 48).
--spec-ngram-map-k4v-min-hits N	Numero minimo di match richiesti per ngram-map-k4v (default: 1).

Comandi Rimosso / Deprecati

Parametro Obsoleto	Sostituto Consigliato
--draft, --draft-n, --draft-max	--spec-draft-n-max oppure --spec-ngram-mod-n-max
--draft-min, --draft-n-min	--spec-draft-n-min oppure --spec-ngram-mod-n-min
--spec-ngram-size-n	Opzioni specifiche --spec-ngram-*-size-n o --spec-ngram-mod-n-match
--spec-ngram-size-m	Opzioni specifiche --spec-ngram-*-size-m
--spec-ngram-min-hits	Opzioni specifiche --spec-ngram-*-min-hits

4. Example-Specific Params (Parametri del Server)

Parametro	Descrizione
-lcs, --lookup-cache-static FNAME	Percorso della cache di ricerca statica per la decodifica speculativa senza modello (non aggiornata durante la generazione).
-lcd, --lookup-cache-dynamic FNAME	Percorso della cache di ricerca dinamica (aggiornata in tempo reale durante la generazione del testo).
-ctxcp, --ctx-checkpoints, --swa-checkpoints N	Numero massimo di punti di ripristino (checkpoint) del contesto creati per ogni slot di sessione utente (default: 32).