Report LLM su Server Lenovo

Framework metodologico per sistemi conversazionali enterprise

Una collaborazione Memori, Lenovo e Araneum e i loro team

a cura di:

Fabio Lecca, Matteo Mastranza, Francesco Piccolo

30/06/2025

Prefazione

Nel contesto attuale, riteniamo che sfruttare in modo consapevole gli strumenti di Intelligenza Artificiale disponibili sia ormai una necessità operativa. Le soluzioni open source, in particolare, offrono vantaggi strategici significativi: consentono di evitare la dipendenza da fornitori esterni, ridurre i costi variabili, utilizzare infrastrutture proprietarie, preservare la riservatezza dei dati sensibili ed evitare interruzioni di servizio non gestibili.

Muovendoci lungo il percorso tracciato nel primo report e facendo nuovamente leva sulla piattaforma Aisuru, che ha confermato la sua efficacia nel semplificare e potenziare l'interazione con i LLM, questo secondo studio si è focalizzato sull'applicazione concreta delle evidenze raccolte. In particolare, sono stati impiegati hardware più performanti, adottati modelli linguistici di ultima generazione e implementate strategie di gestione dei prompt più mirate ed efficienti.

Indice

1. Introduzione

Obiettivi del test

Anche in questo secondo ciclo di test, l'obiettivo principale è valutare l'efficacia di modelli di linguaggio di grandi dimensioni (LLM) open source eseguiti localmente su infrastrutture proprietarie. Questa scelta risponde a due esigenze fondamentali: da un lato, garantire una maggiore tutela dei dati sensibili, evitando l'esposizione a servizi cloud di terze parti; dall'altro, ridurre i costi variabili legati all'utilizzo di API commerciali.

Come nel primo studio, l'analisi si è focalizzata su due aspetti principali:

1. Qualità delle risposte: misurata in termini di pertinenza semantica, consistenza logica e accuratezza dei contenuti generati dai modelli.
2. Misura del carico concorrente: intesa come valutazione della capacità di rispondere a richieste simultanee da parte di utenti concorrenti, senza degradazione significativa delle performance.

Sulla base delle evidenze raccolte nel precedente report, è stato condotto un nuovo ciclo di test applicando le ottimizzazioni individuate, sia sul piano hardware che metodologico.

In particolare, i test sono stati eseguiti su una macchina più potente, in grado di sostenere un carico operativo più elevato. Inoltre, per affrontare il problema riscontrato nella qualità delle risposte su prompt troppo estesi, è stato adottato un approccio più efficiente basato sul function calling. Grazie alle "funzioni avanzate" della piattaforma Aisuru, gli agenti interrogano fonti esterne solo quando necessario, generando prompt più snelli e mirati. Questo riduce la mole informativa e migliora la coerenza delle risposte, diminuendo la possibilità di avere allucinazioni.

La metodologia qualitativa è rimasta invariata rispetto al primo studio, ma è stata applicata in un contesto differente: agenti specializzati su prodotti Lenovo. I test prestazionali, invece, sono stati replicati con lo stesso schema per consentire confronti diretti, ma con modelli più recenti.

2. Server Lenovo

Specifiche Hardware

Il server **Lenovo ThinkSystem SR675 V3** impiegato per l'esecuzione dei test è configurato con componenti hardware di fascia alta, pensati per sostenere carichi di lavoro intensivi legati all'elaborazione di modelli di linguaggio di grandi dimensioni (LLM). Di seguito una panoramica delle principali caratteristiche tecniche.

- **Processore**: il sistema è equipaggiato con 2× AMD EPYC 9535 64-Core Processor, per un totale di 128 core fisici e 256 thread. Ogni CPU opera a una frequenza base di 2.3 GHz, con supporto a boost clock fino a 4.3 GHz, ed è basata sull'architettura x86_64.
- **Memoria RAM**: il sistema dispone di 1.472 GB di RAM DDR5 ECC con una frequenza di 6400 MT/s, distribuiti su 23 moduli da 64 GB ciascuno.
- **GPU**: 2× NVIDIA H200 NVL, ognuna con 120 GB di memoria HBM3 e supporto completo a CUDA 12.8. Queste GPU, di ultima generazione, sono progettate per carichi AI ad altissime prestazioni e operano con un TDP massimo di 600 W per scheda.
- **Storage**: il sistema è dotato di più dispositivi di archiviazione NVMe ad alte prestazioni. Lo spazio di archiviazione complessivo supera i 22 TB, distribuito su diversi dispositivi, tra cui volumi RAID e partizioni dedicate al sistema operativo.
- **Sistema operativo**: il server esegue Ubuntu 22.04.5 LTS a 64 bit, una distribuzione Linux stabile e compatibile con ambienti HPC e AI.

3. Modelli usati per i test

Nello studio sono stati analizzati diversi LLM open source per identificarne l'efficacia nel nostro caso di studio

LLM usati

Modello	Azienda	Context window (Tokens)	Memoria occupata (approx)
Qwen3-32B	Qwen	128k*	30 GB
Qwen3-30B-A3B	Qwen	128k*	28 GB
Qwen3-14B	Qwen	128k*	27.5 GB
Qwen3-8B	Qwen	128k*	7.64 GB
Qwen3-4B	Qwen	128k*	3.8 GB
Qwen3-1.7B	Qwen	32k	1.6GB
Qwen3-0.6B	Qwen	32k	523MB
granite-3.1-8b-instruct	ibm-granite	32k*	7.61GB
granite-3.3-8b-instruct	ibm-granite	128k	7.61GB
granite-3.3-2b-instruct	ibm-granite	128k	2.36GB
gemma-3-27b-it	google	128k	17GB
Llama-3.1-8B-Instruct	meta-llama	128k	4.5GB
Llama-3.3-70B-Instruct	meta-llama	128k	43GB

Llama-3.2-3B	meta-llama	128k	2GB

*I modelli Qwen3 supportano nativamente una context length di 32k ma arrivano a 128k token utilizzando la libreria YaRN supportata da vLLM.

Proprietà dei modelli

Questi modelli hanno delle proprietà che li caratterizzano:

- **Dimensione e memoria**: modelli più grandi offrono prestazioni migliori ma richiedono molta più memoria (fino a 30 GB). I modelli più piccoli sono meno esigenti ma adatti a compiti più semplici.
- **Finestra di contesto**: maggiore è la finestra di contesto, più le richieste possono essere articolate. Tuttavia, l'elaborazione di input molto lunghi può ridurre la qualità delle risposte.

Nonostante siano disponibili modelli open source (es. DeepSeek, Minimax M1, …), con un numero di parametri superiore, e quindi teoricamente anche più potenti, le risorse hardware a disposizione hanno rappresentato un limite pratico. Di conseguenza, Llama-3.3-70B-Instruct ha rappresentato il modello più grande testabile in modo stabile per garantire un livello adeguato di concorrenza nell'ambiente di esecuzione previsto.

4. Agenti su Aisuru

All'interno della piattaforma Aisuru, è stato sviluppato un sistema composto da più agenti, ognuno specializzato in una specifica macro-categoria. Questi agenti sono progettati per rispondere con precisione a richieste relative al proprio ambito di competenza, grazie a una configurazione iniziale definita tramite prompt dedicati.

Tutti gli agenti sono coordinati da una struttura centrale chiamata Board of Experts (BoE), supervisionata da un agente "presidente" denominato Claude L - NOVA. Questo ha il compito di ricevere le domande da parte dell'utente e instradarle verso l'agente più adatto a gestire quel tipo di richiesta, migliorando pertinenza e qualità del risultato finale.

Descrizione Agenti

- **Lenovo Assistant**: Lenovo Assistant è l'esperto utile per:
  - Dare informazioni generiche sul brand Lenovo;
  - Assistere l'utente rispetto a clienti e progetti passati di cui ha già parlato;
  - Aiutare l'utente a creare una mail per il cliente.
- **Consumer**: è l'esperto dei prodotti Consumer di Lenovo, che includono: Notebook Consumer (IdeaPad, Yoga, Legion, LOQ); Desktop Consumer and All in One Consumer (IdeaCentre); Monitor consumer e Monitor Gaming.
  Deve rispondere lui se l'utente:
  - A inizio conversazione ha deciso di parlare di "Consumer" e non ha chiesto esplicitamente di parlare di altri argomenti.
  - Chiede il Part Number (P/N o PN) di un prodotto consumer.
  - Fa domande sui prodotti consumer.

- **Servizi e Garanzie**: è esperto di:
    - Garanzie Lenovo
    - Servizi Lenovo (come x, y, z)
    - Software (come x, y, z)
- **Commercial**: è l'esperto dei prodotti commercial Lenovo, che includono: Notebook Essential: Serie V; Notebook SMB: Thinkbook e Thinkpad E; Notebook Classic. ThinkPad L,T,X e Z; Workstation: Mobile e Desktop Workstation; Desktop: M70, M75, Neo 30, Neo 50; All-in-One: V100, M90a, Neo 50a; Monitor: C series LCD, E series LCD, M series LCD, P series LCD, S series LCD, T series LCD, TIO series LCD; Smart Office: Smart Office; Think Client; Education: 100w, 300w, 500w , 13w Yoga. Deve rispondere lui se l'utente:
    - A inizio conversazione ha deciso di parlare di "Commercial" e non ha chiesto esplicitamente di parlare di altri argomenti.
    - Chiede il Part Number (P/N io PN) di un prodotto commercial.
    - Fa domande sui prodotti commercial.
- **Motorola**: è esperto di Motorola e della sua gamma di prodotti
    - Assiste l'utente rispetto a clienti e progetti passati di cui ha già parlato;
    - Aiuta l'utente a trovare l'offerta commerciale corretta per le esigenze del suo cliente
    - Aiuta l'utente a creare una mail per il cliente.
      La gamma di prodotti Motorola è composta da:
    - razr: è la gamma top premium dei prodotti motorola
    - edge: è la gamma media-alta
    - moto g: è la gamma media
    - moto e: è la gamma entry level
    - thinkphone: è la gamma business
      Deve rispondere lui se l'utente chiede informazioni su smartphone
- **Utility LBP**: crea file Excel e deve rispondere solamente quando l'utente chiede esplicitamente si creare il file excel (per LBP).

Avatar	Nome	Descrizione
	Lenovo Assistant	Lenovo Assistant è l'esperto utile per: - Dare informazioni generiche sul brand Lenovo; - Assistere l'utente rispetto a clienti e progetti passati di cui ha già parlato; - Aiutare l'utente a creare una mail per il cliente.
	Commercial	Commercial è l'esperto dei prodotti commercial Lenovo, che includono: 1) Notebook Essential: Serie V 2) Notebook SMB: Thinkbook e Thinkpad E 3) Notebook Classic. ThinkPad L,T,X e Z 4) Workstation: Mobile e Desktop Workstation 5) Desktop: M70, M75, Neo 30, Neo 50. 6) Espandi
	Consumer	Consumer è l'esperto dei prodotti Consumer di Lenovo, che includono: - Notebook Consumer (IdeaPad, Yoga, Legion, LOQ) - Desktop Consumer and All in One Consumer (IdeaCentre) - Monitor consumer e Monitor Gaming Deve rispondere lui se l'utente: -A inizio Espandi
	Servizi e Garanzie	Servizi e Garanzie è esperto di: - Garanzie Lenovo - Servizi Lenovo (come x, y, z) - Software (come x, y, z)
	Motorola	MOTOROLA è esperto di Motorola e della sua gamma di prodotti; - Assiste l'utente rispetto a clienti e progetti passati di cui ha già parlato; - Aiuta l'utente a trovare l'offerta commerciale corretta per le esigenze del suo cliente - Aiuta l'utente a creare una mail per il Espandi
	Utility LBP	Utility LBP crea file Excel E DEVE RISPONDERE SOLAMENTE QUANDO L'UTENTE CHIEDE ESPLICITAMENTE SI CREARE IL FILE EXCEL (per LBP).

Funzioni avanzate

Ogni agente, oltre al proprio prompt istruttivo, è abilitato a chiamare delle funzioni avanzate nel momento in cui le informazioni richieste dall'utente non sono disponibili nella conoscenza pre-addestrata del modello. Queste funzioni permettono di consultare dati aggiornati, come schede prodotto o elenchi di caratteristiche tecniche, migliorando l'affidabilità delle risposte.

Le funzioni vengono attivate solo quando ritenute necessarie, e si basano su webhook che puntano a pagine HTML formattate in modo tabellare, contenenti tutte le informazioni utili all'elaborazione. Ad esempio, se un utente chiede un consiglio sull'acquisto di un portatile per uso ufficio, NOVA inoltra la richiesta all'agente Consumer, che può decidere di attivare la funzione "NOTEBOOK_CONSUMER" per accedere a tutti i dati relativi alla gamma notebook consumer Lenovo.

Attualmente, gli agenti abilitati all'utilizzo di queste funzioni per il recupero di informazioni prodotto sono: MOTOROLA, CONSUMER e COMMERCIAL. Questo approccio consente alla piattaforma di mantenere le risposte aggiornate, accurate e coerenti con l'offerta reale, pur sfruttando le capacità di ragionamento e linguaggio naturale del modello.

Funzioni avanzate dell'agente Consumer

Nome	Descrizione
DESKTOP_AND_AIO_CONSUMER	Questa funzione contiene dati relativi a Desktop e All-in-One Consumer. Utilizza questa funzione quando devi ottenere informazioni su prezzi, promo, specifiche tecniche e simili.
MONITOR_CONSUMER	Questa funzione contiene dati relativi a Monitor Consumer e Monitor Gaming. Utilizza questa funzione quando devi ottenere informazioni su prezzi, promo, specifiche tecniche e simili.
NOTEBOOK_CONSUMER	La funzione contiene dati relativi ai notebook consumer: IdeaPad, Yoga, Legion e LOQ. Utilizza questa funzione quando è necessario ottenere dati su prezzi, promo, specifiche tecniche e simili.

Questo approccio di partizionamento dei dati consente di ottimizzare i token consumati, ma può avere dei limiti in caso di richieste troppo generiche dell'utente, ad es. "Indicami il prodotto meno costoso", perchè i modelli LLM tendono ad eseguire una sola funzione per rispondere alla domanda dell'utente, quindi in caso di informazioni parziali potrebbero dare risposte incomplete in quanto basate sui dati presenti in un singolo dataset. Risulta quindi importante definire correttamente il prompt per chiedere all'utente le informazioni necessarie per selezionare la tipologia di prodotto di interesse.

5. Valutazione qualitativa delle risposte

Conversazioni da valutare

Per valutare l'efficacia e la qualità delle risposte fornite dagli agenti sulla piattaforma Aisuru, Lenovo ha messo a disposizione un file Excel contenente 12 conversazioni simulate,

ognuna composta da una serie di domande coerenti tra loro, appartenenti a un unico filo logico. Le domande riflettono richieste comuni da parte degli utenti, come suggerimenti su prodotti specifici, indicazioni in base al budget disponibile, oppure esigenze legate a prestazioni o casi d'uso particolari.

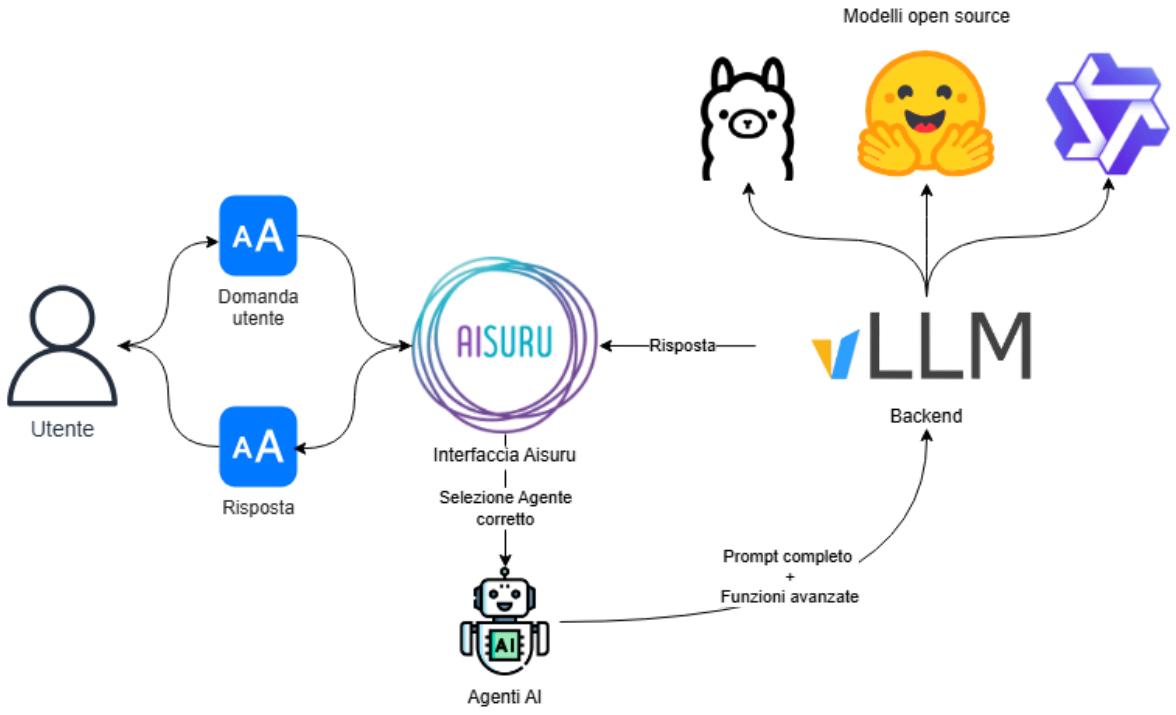
Per ciascuna domanda, il file Excel includeva anche:

- l'agente della BoE che avrebbe dovuto essere coinvolto per gestirla correttamente;
- una risposta attesa di riferimento, utile per valutare l'adeguatezza della risposta effettivamente generata dal sistema
- la registrazione della fase di reasoning del modello (se prevista dal modello), per poter valutare anche la correttezza del ragionamento che ha portato alla risposta ottenuta

Durante questa fase di test, la piattaforma Aisuru è stata configurata per utilizzare un diverso backend LLM: al posto dei modelli commerciali (come Claude o OpenAI), è stato attivato il collegamento al nostro server Lenovo su cui era installata una versione di vLLM con modelli open source. Questo ha permesso di valutare la performance dei modelli open in un contesto realistico, mantenendo invariata l'architettura della BoE e la logica di smistamento tra gli agenti.

Le risposte sono state generate automaticamente tramite uno script che, a partire dalle domande presenti nel file, le inviava al sistema e salvava le risposte ricevute. Queste sono state poi archiviate in un nuovo file, che ha costituito la base per l'analisi qualitativa.

Architettura



Risultati

I risultati ottenuti dalla valutazione qualitativa confermano le aspettative, ovvero i modelli Qwen3 a partire dalla versione 8B iniziano ad essere sufficienti per il compito richiesto, e ovviamente all'aumentare dei parametri i risultati diventano sempre più corretti. Per quanto riguarda i modelli più piccoli, apparentemente le risposte possono sembrare corrette, ma andando a leggere i contenuti effettivi, essi generano informazioni sulla base della loro conoscenza pregressa, a volte senza sfruttare i dati prelevati dalle funzioni chiamate. Dalle conversazioni salvate, si nota come in alcuni casi il modello genera Part Number di modelli che non esistono o con specifiche che non sono coerenti con la realtà.
In particolare si raccomanda il modello 14B come base minima per l'utilizzo con agenti che devono utilizzare tool. Tale modello utilizza 27.5 Gb di RAM della GPU, quindi può essere istanziato più volte oppure si può lasciare la gestione completa della memoria delle GPU a vLLM, per garantire l'efficienza del parallelismo.

6. Test di prestazioni

L'obiettivo principale del test è identificare il giusto compromesso tra tre fattori fondamentali:

- **Qualità delle risposte**: ovvero la capacità del modello di fornire output coerenti, rilevanti e precisi rispetto alla richiesta
- **Tempo di risposta**: parametro cruciale in applicazioni interattive o real-time, che misura la reattività percepita dal punto di vista dell'utente
- **Numero di richieste gestibili nell'unità di tempo**: indicatore diretto della scalabilità del sistema, fondamentale per mantenere performance accettabili anche in scenari ad alto traffico.

Questi tre elementi sono correlati tra loro: migliorare la qualità richiede modelli più complessi e pesanti, che però aumentano i tempi di risposta e riducono il throughput complessivo. Al contrario, l'utilizzo di modelli più leggeri migliora scalabilità e latenza, ma può compromettere la qualità delle risposte. L'equilibrio tra questi fattori è quindi essenziale per garantire un'esperienza d'uso ottimale all'interno di un contesto come quello analizzato.

Nell'ambito della valutazione delle capacità dei modelli linguistici (LLM) a nostra disposizione, abbiamo condotto un'analisi approfondita delle loro performance utilizzando due strumenti principali: Apache JMeter e un sistema di benchmark interno a vLLM.

Test Apache JMeter

In quest'ottica, il test di performance è stato progettato con carichi progressivi, fino a raggiungere un punto di effettivo stress dell'infrastruttura, identificato nel carico di 200 utenti simultanei. In questa configurazione, sono stati condotti test comparativi utilizzando i diversi modelli Qwen3, variando il numero di parametri per valutare l'impatto su qualità, latenza e throughput.

Il prompt utilizzato ha una lunghezza di 50.000 byte e contiene al suo interno anche la definizione di tool, in modo da simulare una richiesta tipica proveniente da Aisuru, quindi il

throughput corrisponde ad un utilizzo con tale piattaforma più possibile simile alle applicazioni reali in produzione.

Modello	Memoria occupata	Tempo	Throughput/s	Avg msec	Min msec	Max msec
Qwen3 32b	30 GB	06:03	5.5	32932	10540	62336
Qwen3 14b	27.52 GB	03:02	11.0	16589	6581	30920
Qwen3 8b	7.64 GB	02:30	13.3	14101	5449	23780
Qwen3 4b	3.8 GB	03:01	11.0	16767	4978	29535
Qwen3 0.6b	0.5 GB	01:08	29.6	6244	1627	10682

7. Benchmark vLLM

Questo capitolo fornisce un'analisi quantitativa delle performance dei modelli testati, utilizzando un dataset di circa **53k conversazioni ShareGPT** derivate da un dataset iniziale di ~100 k interazioni presente su Hugging Face: https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/blob/main/ShareGPT_V3_unfiltered_cleaned_split.json

vLLM

vLLM è una libreria open-source progettata per l'inferenza e il servizio di modelli di linguaggio di grandi dimensioni (LLM) in modo efficiente e ad alte prestazioni. Con il comando *serve* si possono istanziare i LLM installati per interrogarli e valutarne le prestazioni.

Metodologia di benchmark

I test sono stati condotti utilizzando input standardizzati e un limite massimo di token per la risposta (**Output-len**) impostato a 128 token. Per ogni modello, sono stati raccolti i seguenti indicatori di performance:

- **throughput (requests/s)**: il numero di richieste completate al secondo, che rappresenta un indicatore diretto della capacità del modello di gestire carichi multiutente.
- **total tokens/s**: il totale dei token processati al secondo, inclusi input ed elaborazione interna.
- **output tokens/s**: il numero di token generati dal modello al secondo, un indicatore dell'efficienza del modello nella produzione di risposte.

Comando utilizzato per far partire i test:
`./vllm/benchmarks/benchmark_throughput.py --dataset ShareGPT_V3_unfiltered_cleaned_split.json --output-len 128 --model <modello> --tensor-parallel-size 2 --gpu-memory-utilization 0.95`

Modello	Requests/s	Total tokens/s	Output tokens/s
ibm-granite/granite-3.1-8b-instruct	85,46	32.596,04	10.939,31
ibm-granite/granite-3.3-2b-instruct	115,08	43.893,51	14.730,41
ibm-granite/granite-3.3-8b-instruct	82,42	31.435,21	10.549,47
Qwen/Qwen3-0.6B	150,43	53.755,25	19.255,58
Qwen/Qwen3-1.7B	134,34	48.003,86	17.195,38
Qwen/Qwen3-4B	108,78	38.870,13	13.923,60
Qwen/Qwen3-8B	90,91	32.483,51	11.635,86
Qwen/Qwen3-14B	65,50	23.404,51	8.383,69
Qwen/Qwen3-32B	35,24	12.591,11	4.510,10
Qwen/Qwen3-30B-A3B	75,35	26.925,00	9.466,76
google/gemma-3-27b-it	28,24	10.713,14	3.691,88
meta-llama/Llama-3.2-3B	129,53	45.638,10	16.580,13
meta-llama/Llama-3.3-70B-Instruct	18,77	6.612,60	2.402,33
meta-llama/Llama-3.1-8B-Instruct	88,73	31.263,15	11.357,77

Comparazione con la precedente infrastruttura

Per valutare i benefici dell'aggiornamento dell'infrastruttura, sono stati comparati i benchmark tra il server precedente e il nuovo server. Sebbene non siano stati utilizzati esattamente gli stessi modelli nelle due misurazioni, è stato scelto per ciascun caso il corrispettivo aggiornato più recente (es. da Llama 3.1 a 3.3, da Qwen2.5 a Qwen3, ecc.), in quanto sarebbe stato poco significativo ripetere i test con versioni obsolete.
Inoltre per fare una comparazione efficace, sono stati condotti anche altri benchmark con lunghezza di output pari a quella dei test del server precedente.

512 output-len	Modello	Req/s	Total tokens/s	Output tokens/s	Miglioramento
Nuovo	Qwen/Qwen3-32B	13,57	10.056,72	6.945,64	3,02
Precedente	Qwen/Qwen2.5-32B-Instruct-AWQ	4,50	3.332,53	2.301,60	
Nuovo	google/gemma-3-27b-it	8,58	6.482,23	4.393,39	1,95
Precedente	google/gemma-2-27b-it	4,41	3.321,08	2.255,98	
Nuovo	ibm-granite/granite-3.1-8b-instruct	27,29	20.885,14	13.970,49	3,10
Precedente	ibm-granite/granite-3.1-8b-instruct	8,80	6.732,69	4.503,63	
128 output-len					
Nuovo	meta-llama/Llama-3.3-70B-Instruct	18,77	6.612,60	2.402,33	2,85
Precedente	casperhansen/llama-3.3-70b-instruct-awq	6,58	2.316,96	841,74	
Nuovo	meta-llama/Llama-3.1-8B-Instruct	88,73	31.263,15	11.357,77	3,15
Precedente	meta-llama/Llama-3.1-8B-Instruct	28,17	9.925,17	3.605,77	
Nuovo	Qwen/Qwen3-14B	65,50	23.404,51	8.383,69	3,82
Precedente	Qwen/Qwen2.5-14B-Instruct-AWQ	17,13	6.122,48	2.193,12	

Considerazioni

Il confronto ha evidenziato un miglioramento significativo e consistente su tutta la linea. In media, il nuovo server ha mostrato una performance circa 3 volte superiore rispetto al precedente, in tutte e tre le metriche analizzate.
In particolare, si osservano:

- Incrementi sostanziali nei modelli più piccoli di 8 miliardi di parametri, che raggiungono fino a **3.1–3.2×** di miglioramento.
- Un miglioramento anche sui modelli più grandi (es. Llama 70B e Qwen 32B), dove le nuove versioni aggiornate, accoppiate al nuovo server, hanno mostrato un throughput decisamente più alto, con aumenti superiori al **2.8×**.
- Il miglioramento più grande lo abbiamo riscontrato nel modello con 14 miliardi di parametri, dove c'è stato un aumento del **3.8x**.

Nel complesso, questi risultati confermano che il passaggio al nuovo server ha avuto un impatto diretto e quantificabile in termini di prestazioni, garantendo maggiore efficienza nell'esecuzione dei modelli di inferenza e una maggiore scalabilità per carichi futuri.

8. Conclusioni

Lo studio corrente è stato completato a soli 6 mesi di distanza dal precedente, e nel frattempo abbiamo avuto a disposizione un'infrastruttura notevolmente più performante dal punto di vista hardware, nonché una generazione successiva per tutti i modelli open source che erano stati testati a gennaio 2025. Questi due fattori lasciano ben sperare sul fatto che le soluzioni LLM on premise saranno via via sempre più utilizzabili in ambiti lavorativi dove la privacy e la disponibilità dei dati aziendali sia un fattore determinante per la scelta di tale

soluzione. La nuova infrastruttura ha mostrato un incremento di performance di circa 4x rispetto alla precedente, consentendo anche l'utilizzo di modelli con maggior numero di parametri. Sono stati identificati dei modelli open source che si integrano correttamente con la piattaforma AISuru e consentono anche un alto grado di parallelismo delle richieste, in modo da consentire a più utenti aziendali simultanei di interagire correttamente con gli agenti.