

SYSTEM-LEISTUNGSTEST

Lokale LLM-Inferenz Ubuntu vs. Windows

AMD Ryzen Threadripper PRO 9955WX · NVIDIA GeForce RTX 5090 32 GB · ca. 256 GB RAM

Datenbasis

Ubuntu-Referenzlauf Docker/CUDA, 19 h 18 min.

Windows-Referenzlauf: Native CUDA-Ausführung, 38 h 59 min.

Beide Referenzläufe: sechs Modelle, 10 Wiederholungen, Prompt bis 32K, Generation bis 1024 Tokens, Long Context bis 262K.

1. Zusammenfassung

Gesamtbewertung

Das getestete System erreicht unter beiden Betriebssystemen High-End-Leistung für lokale LLM-Inferenz. Die RTX 5090 liefert bei Modellen, die vollständig in 32 GB VRAM passen, sehr hohe Einzel- und Serverleistung. Der Threadripper und der große Arbeitsspeicher ermöglichen zusätzlich die lokale Ausführung von 119B-/122B-Modellen auf der CPU. Die thermische Stabilität ist unter 30-minütiger Dauerlast ausgezeichnet; es wurde kein relevanter Leistungseinbruch beobachtet.

- Single-Stream-GPU: Ubuntu liegt bei den vier vollständig GPU-fähigen Modellen meist ca. 0–4 % vor Windows; die Unterschiede sind insgesamt klein.
- Schnellstes getestetes Generierungsmodell: GLM-4.7-Flash mit 257,7 tok/s unter Ubuntu und 253,7 tok/s unter Windows.
- Stärkste Prompt-Verarbeitung: Qwen3.5-9B mit 10.908 tok/s (Ubuntu) bzw. 10.498 tok/s (Windows) bei 32K Prompt.
- Long Context: Gemma 4 12B hält die Leistung am stabilsten und erreicht bei 262K noch 107,8 tok/s (Ubuntu) bzw. 105,8 tok/s (Windows).
- Serverlast: Ubuntu erreicht bei 32 parallelen Requests 1.409,6 System-tok/s; Windows 1.176,9 System-tok/s. Beide Läufe bleiben vollständig erfolgreich.
- Multi-Tenant: Zwei Modelle gleichzeitig erreichen bei jeweils Concurrency 8 zusammen ca. 954,5 tok/s unter Ubuntu und 1.330,4 tok/s unter Windows.
- Dauerlast: GPU-Maximaltemperaturen liegen je nach Modell zwischen 61 und 71 °C; kein relevanter thermischer Durchsatzverlust.
- Kapazitätsgrenze: 119B-/122B-GGUFs mit ca. 69–71 GiB passen nicht vollständig in 32 GB VRAM. CPU-only bleibt jedoch nutzbar; erzwungener Windows-GPU-Overload ist deutlich langsamer als CPU-only.

2. Testsystem und Vergleichbarkeit

Merkmal	Ubuntu	Windows
Ausführung	Docker / Ubuntu 24.04 Runtime	Windows 11 nativ
CPU	Threadripper PRO 9955WX, 16C/32T	Threadripper PRO 9955WX, 16C/32T
Arbeitsspeicher	251.1 GiB	255.4 GiB
GPU	GeForce RTX 5090, 32 GB	GeForce RTX 5090, 32 GB
GPU Power Limit	600 W	600 W
NVIDIA-Treiber	610.57.04	617.14
CUDA-/Runtime	CUDA 13.2.1 Container	CUDA 13.4 Build
llama.cpp Build	64a155d24 / 10808	25747b08e / 11284
llmbench	1.5.0	1.5.0
Referenzlauf	19,3 h	39,0 h

Hinweis zur Interpretation

Die Hardware und die Benchmarkparameter sind praktisch identisch, die Laufzeitumgebungen und llama.cpp-Builds unterscheiden sich jedoch

3. Testumfang und Modelle

Modell	GGUF-Größe	Rolle im Test
Gemma 4 12B	6,6 GiB	GPU + CPU
GLM-4.7-Flash	16,9 GiB	GPU + CPU
Qwen3.5-9B	5,3 GiB	GPU + CPU
Qwen3.5-72B	16,2 GiB	GPU + CPU
Mistral Small 4 119B	68,7 GiB	CPU; Full-GPU >32 GB VRAM
Qwen3.5-122B-A10B	71,3 GiB	CPU; Full-GPU >32 GB VRAM

Gemeinsame Referenzparameter: **10 Wiederholungen, Batch 2048, UBatch 512, Flash Attention auto, KV-Cache F16, Promptgrößen 512/4K/8K/16K/32K, Generation 128/512/1024 und Kontextstufen 0/8K/32K/65K/131K/262K.**

4. Ubuntu-Ergebnisse

Der Ubuntu-Referenzlauf wurde im Docker-Container mit CUDA 13.2.1 ausgeführt. Die CPU lief im Performance-Governor. Der Lauf dauerte rund 19,3 Stunden.

4.1 GPU-Einzelmodellleistung

Modell	tg128	tg1024	PP 32K
Gemma 4 12B	159,7	158,4	7713,5
GLM-4.7-Flash	257,7	256,0	3563,3
Qwen3.5-9B	229,3	230,0	10908,2
Qwen3.8-27B	78,7	78,9	3416,5

- GLM-4.7-Flash erzielt die höchste Generierungsleistung mit 257,7 tok/s bei tg128.
- Qwen3.5-9B ist bei der Prompt-Verarbeitung klar führend und erreicht 10.908 tok/s bei 32K Prompt.
- Qwen3.8-27B liegt bei rund 78,7 tok/s und zeigt damit die erwartete Skalierung eines größeren dichten Modells.

4.2 CPU-Generierungsleistung

Modell	CPU tg128	CPU tg1024
Gemma 4 12B	9,9	9,9
GLM-4.7-Flash	30,5	27,2
Qwen3.5-9B	12,6	13,6
Qwen3.8-27B	4,4	4,7
Mistral Small 4 119B	17,4	17,0
Qwen3.5-122B-A10B	8,3	7,9

Messhinweis Ubuntu CPU

Die Ubuntu-CPU-Prompt-Processing-Messungen zeigten während der Prompt-Phase deutliche GPU-Aktivität und werden deshalb nicht als reine CPU-Promptwerte in den Betriebssystemvergleich übernommen. Die CPU-Generierungsphasen zeigten dagegen überwiegend nur geringe GPU-Aktivität und werden als Leistungsindikator aufgeführt.

4.3 Long-Context-Leistung der GPU

Modell	0K	32K	65K	131K	262K
Gemma 4 12B	159,7	146,0	139,0	126,8	107,8
GLM-4.7-Flash	258,5	169,5	133,3	98,4	64,9
Qwen3.5-9B	229,5	198,3	174,9	141,1	102,0
Qwen3.8-27B	78,8	71,2	65,2	55,7	—

- Gemma 4 12B verliert von 0K bis 262K nur rund 32,5 % und ist damit das stabilste Modell bei extrem langem Kontext.
- Qwen3.5-9B erreicht bei 262K noch 102,0 tok/s.
- GLM-4.7-Flash fällt von 258,5 auf 64,9 tok/s und reagiert am stärksten auf sehr große Kontexte.
- Qwen3.8-27B wurde bis 131K vollständig gemessen (55,7 tok/s); die 262K-Stufe wurde nicht abgeschlossen.

4.4 30-Minuten-Dauerlast

Modell	GPU tok/s	TPS-Verlust	Ø W	Peak W	Max. Temp.
Gemma 4 12B	111,9	0,5 %	488,3	539,2	68 °C
GLM-4.7-Flash	164,6	-0,3 %	367,6	397,7	62 °C
Qwen3.5-9B	140,1	-1,8 %	460,4	541,9	68 °C
Qwen3.8-27B	55,8	0,2 %	528,5	604,3	71 °C

Dauerlastbewertung Ubuntu

Keines der vier GPU-fähigen Modelle zeigt relevantes thermisches Throttling. Selbst Qwen3.8-27B erreicht im 30-Minuten-Lauf im Mittel 528,5 W, bis zu 604,3 W und maximal 71 °C bei praktisch konstantem Durchsatz.

4.5 Server- und Parallelitätsleistung

Concurrency	System tok/s	tok/s je Request	TTFT P50	TTFT P95
1	150,9	150,9	41 ms	43 ms
2	275,1	137,6	68 ms	71 ms
4	455,4	113,9	71 ms	99 ms
8	666,7	85,3	226 ms	239 ms
16	1051,6	65,8	425 ms	440 ms
32	1409,6	44,1	707 ms	857 ms

Bei Concurrency 32 wurden 1.409,6 System-tok/s erreicht. Die GPU lag im gesamten TTFT-Test im Mittel bei 86.3 % Auslastung und 518.0 W; Spitzenwert 602.5 W, Maximaltemperatur 72 °C.

4.6 Multi-Tenant und Kontextkapazität

- Qwen3,5-9B und Gemma 4 12B liefen gleichzeitig, Bei jeweils Concurrency 8 wurden zusammen 954,5 tok/s erreicht,
- Die GPU erreichte dabei 100 % Spitzenlast, rund 608 W Spitze und maximal 70 °C.
- Der progressive Kontext-Test bestätigte Gemma 4 Full-GPU stabil bis mindestens 130.012 Prompt-Tokens.

5. Windows-Ergebnisse

Der Windows-Referenzlauf wurde nativ unter Windows 11 mit CUDA 13.4 ausgeführt. Der Lauf dauerte rund 39,0 Stunden und nutzte dieselbe Modell- und Testauswahl.

5.1 GPU-Einzelmodellleistung

Modell	tg128	tg1024	PP 32K
Gemma 4 12B	159,7	155,9	7596,4
GLM-4.7-Flash	253,7	248,6	3532,9
Qwen3.5-9B	221,6	223,5	10498,0
Qwen3.8-27B	77,6	78,6	3376,2

5.2 CPU-Generierungsleistung

Modell	CPU tg128	CPU tg1024
Gemma 4 12B	13,3	13,1
GLM-4.7-Flash	32,6	30,2
Qwen3.5-9B	18,2	18,0
Qwen3.8-27B	5,8	5,8
Mistral Small 4 119B	19,7	19,1
Qwen3.5-122B-A10B	10,1	10,1

- Mistral Small 4 119B erreicht trotz rund 68,7 GiB Modellgröße 19,7 tok/s auf der CPU.
- GLM-4.7-Flash ist mit 32,6 tok/s das schnellste CPU-Generierungsmodell.
- Die Windows-CPU-Messungen zeigen im Mittel nur rund 0,3 % GPU-Auslastung und sind damit als echte CPU-only-Werte gut interpretierbar.

5.3 Long-Context-Leistung der GPU

Modell	0K	32K	65K	131K	262K
Gemma 4 12B	156,9	143,1	136,8	124,4	105,8
GLM-4.7-Flash	253,8	173,4	138,0	102,4	67,2
Qwen3.5-9B	220,1	193,6	173,0	139,0	100,9
Qwen3.8-27B	78,2	71,4	65,1	55,8	—

- Gemma 4 12B erreicht bei 262K noch 105,8 tok/s und verliert nur rund 32,6 % gegenüber 0K.
- Qwen3.5-9B erreicht bei 262K 100,9 tok/s.
- GLM-4.7-Flash erreicht bei 262K 67,2 tok/s.
- Qwen3.8-27B liefert bis 131K 55,8 tok/s; die 262K-Stufe wurde innerhalb des Zeitfensters nicht vollständig beendet.

5.4 30-Minuten-Dauerlast

Modell	GPU tok/s	TPS-Verlust	Ø W	Peak W	Max. Temp.
Gemma 4 12B	108,7	-0,0 %	456,3	508,6	66 °C
GLM-4.7-Flash	166,5	-0,2 %	356,0	383,8	61 °C
Qwen3.5-9B	137,9	0,8 %	436,4	496,0	67 °C
Qwen3.8-27B	58,1	0,4 %	526,1	591,3	71 °C

Dauerlastbewertung Windows

Auch unter Windows zeigen alle vier GPU-fähigen Modelle stabile 30-Minuten-Werte. Der höchste gemessene Wert liegt bei 591,3 W und 71 °C (Qwen3.8-27B); der Durchsatzverlust bleibt deutlich unter 1 %.

5.5 Server- und Parallelitätsleistung

Concurrency	System tok/s	tok/s je Request	TTFT P50	TTFT P95
1	148,0	148,2	43 ms	51 ms
2	267,6	134,1	76 ms	90 ms
4	461,2	115,7	80 ms	112 ms
8	673,2	84,6	234 ms	249 ms

16	931,7	58,7	470 ms	478 ms
32	1176,9	37,1	760 ms	893 ms

Bei Concurrency 32 wurden 1.176,9 System-tok/s erreicht. Die GPU lag im TTFT-Test im Mittel bei 81.8 % und 492.0 W; Spitzenwert 604.6 W, Maximaltemperatur 72 °C.

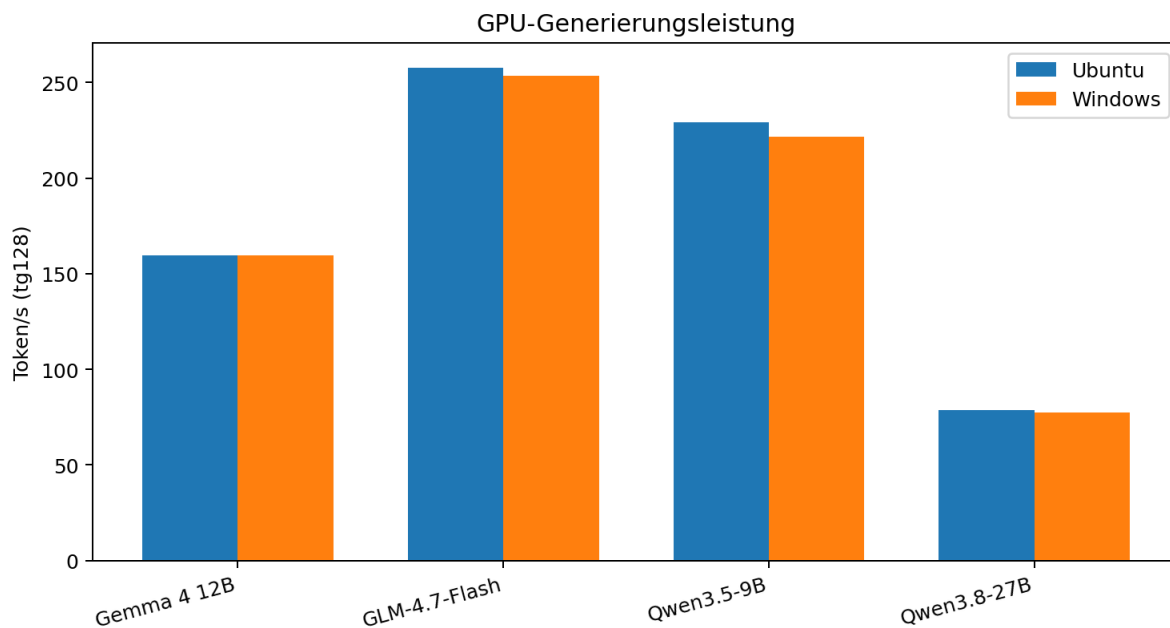
5.6 Multi-Tenant, Kontext und Overload

- Qwen3.5-9B + Gemma 4 12B erreichen bei jeweils Concurrency 8 zusammen 1330,4 tok/s,
- Dabei werden 100 % GPU-Spitzenlast, rund 603 W und maximal 70 °C erreicht.
- Gemma 4 wurde im Kontext-Stresstest stabil bis mindestens 130.012 Prompt-Tokens bestätigt.
- Im separaten Windows-Overload-Test erreichen Mistral 119B nur 1,65 tok/s und Qwen3.5-122B 5,32 tok/s, wenn die deutlich zu großen Modelle über 32 GB VRAM hinaus erzwungen werden.

6. Direkter Ubuntu-/Windows-Vergleich

6.1 GPU-Einzelstream

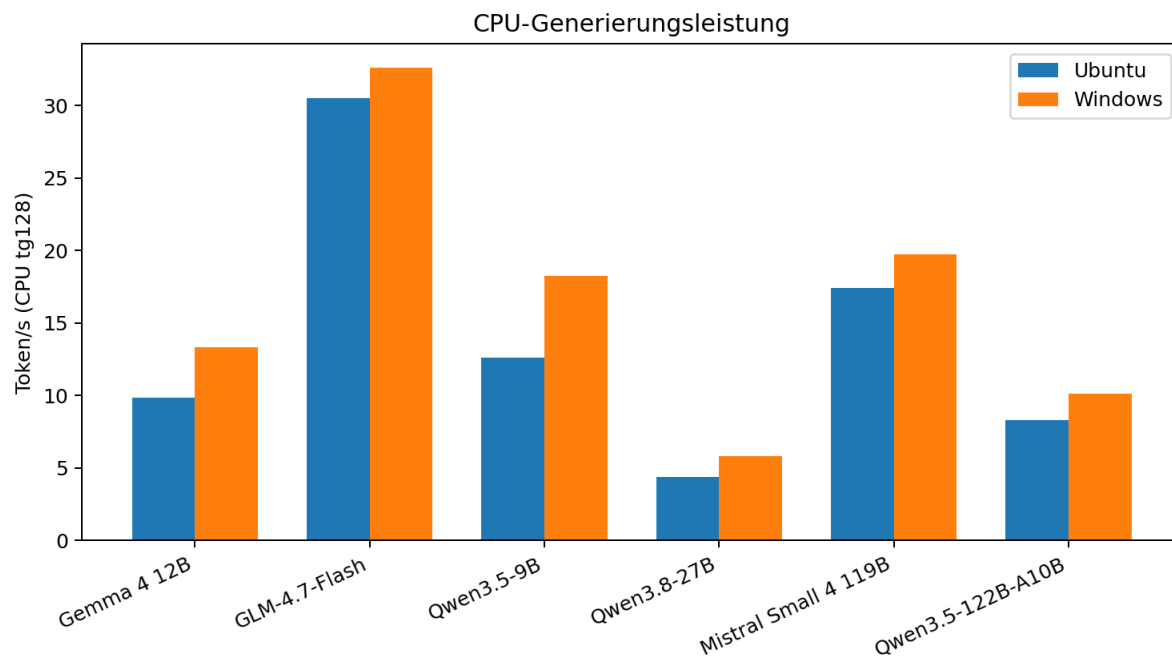
Modell	Ubuntu tg128	Windows tg128	Windows vs. Ubuntu
Gemma 4 12B	159,7	159,7	-0,0 %
GLM-4.7-Flash	257,7	253,7	-1,6 %
Qwen3.5-9B	229,3	221,6	-3,4 %
Qwen3.8-27B	78,7	77,6	-1,4 %



Ergebnis: **Die GPU-Einzelstream-Leistung ist unter beiden Plattformen sehr ähnlich. Ubuntu liegt in diesen Läufen je nach Modell zwischen praktisch gleichauf und rund 3,4 % vor Windows.**

6.2 CPU-Generation

Modell	Ubuntu tg128	Windows tg128	Windows vs. Ubuntu
Gemma 4 12B	9,9	13,3	35,2 %
GLM-4.7-Flash	30,5	32,6	7,0 %
Qwen3.5-9B	12,6	18,2	44,7 %
Qwen3.8-27B	4,4	5,8	33,0 %
Mistral Small 4 119B	17,4	19,7	13,1 %
Qwen3.5-122B-A10B	8,3	10,1	22,1 %



In der reinen Generierungsphase liegt Windows bei allen sechs Modellen vor Ubuntu. Der Abstand reicht von etwa 7 % (GLM) bis rund 45 % (Qwen3.5-9B). Die Ubuntu-Prompt-Processing-CPU-Werte werden wegen der dort beobachteten GPU-Aktivität nicht für einen direkten CPU-Promptvergleich verwendet.

6.3 Long Context

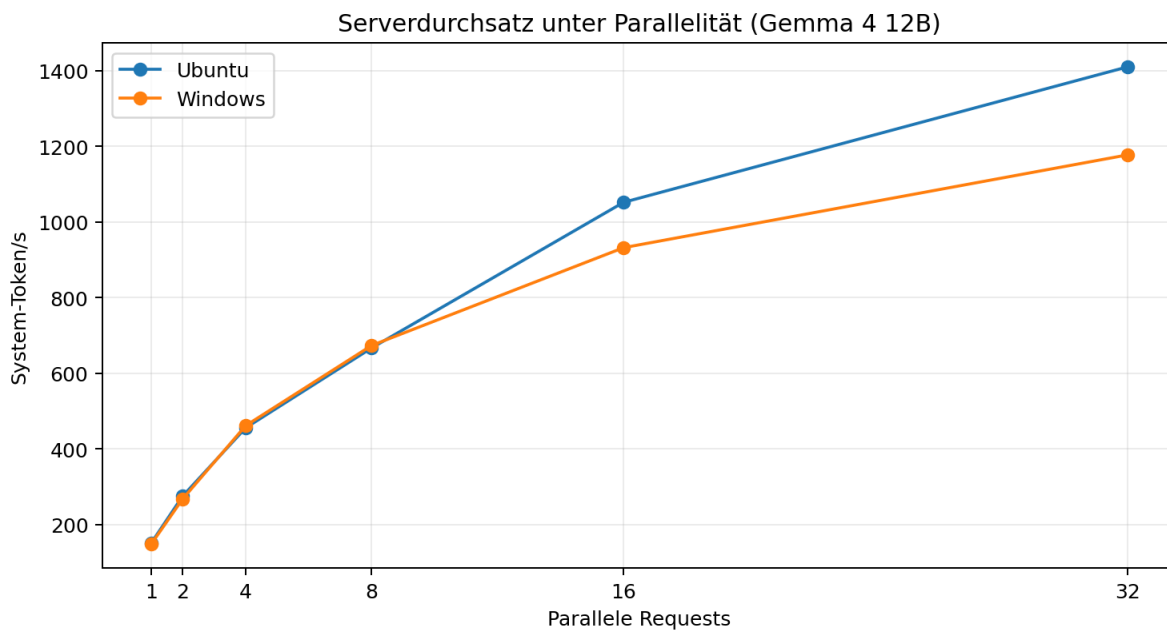
Modell	Ubuntu max. Stufe	tok/s	Windows max. Stufe	tok/s
Gemma 4 12B	262K	107,8	262K	105,8
GLM-4.7-Flash	262K	64,9	262K	67,2
Qwen3.5-9B	262K	102,0	262K	100,9
Qwen3.8-27B	131K	55,7	131K	55,8

Long-Context-Fazit

Die Plattformunterschiede sind klein. Gemma, Qwen3.5-9B und GLM schaffen unter beiden Systemen 262K. Qwen3.8-27B ist bis 131K praktisch identisch (55,73 vs. 55,77 tok/s).

6.4 Serving-Skalierung

Conc.	Ubuntu tok/s	Windows tok/s	Win vs. Ub	Ubuntu P95	Windows P95
1	150,9	148,0	-1,9 %	43 ms	51 ms
2	275,1	267,6	-2,7 %	71 ms	90 ms
4	455,4	461,2	1,3 %	99 ms	112 ms
8	666,7	673,2	1,0 %	239 ms	249 ms
16	1051,6	931,7	-11,4 %	440 ms	478 ms
32	1409,6	1176,9	-16,5 %	857 ms	893 ms



- Bis Concurrency 8 liegen beide Plattformen sehr nah beieinander.
- Bei Concurrency 16 und 32 skaliert der Ubuntu-Lauf stärker: 1.051,6 bzw. 1.409,6 System-tok/s gegenüber 931,7 bzw. 1.176,9 unter Windows.
- Die P95-TTFT bleibt bei 32 parallelen Requests auf beiden Systemen unter 0,9 Sekunden.

6.5 Multi-Tenant

Messpunkt	Ubuntu	Windows
Qwen3.5-9B @ C8	472,2 tok/s	649,4 tok/s
Gemma 4 12B @ C8	482,3 tok/s	681,0 tok/s
Kombiniert @ C8	954,5 tok/s	1330,4 tok/s
GPU Ø Auslastung	91,9 %	ca. 95,0 %
GPU Ø Leistung	523,5 W	ca. 536 W
Max. Temperatur	70 °C	70 °C

Im Multi-Tenant-Szenario liegt der Windows-Lauf bei hoher Parallelität deutlich vorne: 1.330,4 tok/s kombiniert gegenüber 954,5 tok/s unter Ubuntu. Das steht im Gegensatz zum TTFT-Einzelservers-Test, bei dem Ubuntu bei sehr hoher Parallelität stärker skaliert. Die beiden Tests belasten unterschiedliche Serverkonstellationen und sollten daher getrennt bewertet werden.

7. Kapazität, Thermik und praktische Einsatzbewertung

7.1 VRAM- und Modellgrößengrenzen

- Gemma 4 12B, GLM-4.7-Flash, Qwen3.5-9B und Qwen3.8-27B passen vollständig in die RTX 5090 und liefern hohe GPU-Leistung.
- Mistral Small 4 119B (~68,7 GiB) und Qwen3.5-122B (~71,3 GiB) überschreiten die 32-GB-VRAM-Kapazität deutlich.
- Auf der CPU bleiben beide Großmodelle nutzbar: Windows 19,7 tok/s (Mistral 119B) und 10,1 tok/s (Qwen 122B); Ubuntu 17,4 bzw. 8,3 tok/s.
- Der Windows-Overload-Test zeigt, dass erzwungenes Auslagern der Großmodelle über den VRAM hinaus nicht sinnvoll ist: 1,65 tok/s beim Mistral und 5,32 tok/s beim Qwen 122B, also langsamer als CPU-only.

7.2 Thermische Reserve

Thermisches Ergebnis

Unter beiden Betriebssystemen bleibt die RTX 5090 selbst bei 30 Minuten kombinierter CPU/GPU-Dauerlast im Bereich von maximal 61–71 °C, je nach Modell. Die Spitzenleistungsaufnahme liegt in den härtesten Tests bei rund 600 W. Es gibt keinen relevanten Hinweis auf thermisches Throttling.

7.3 Empfohlene Einsatzbereiche

Einsatz	Geeignete Konfiguration	Bewertung
Schnelle Einzelanfragen	GLM-4.7-Flash Full-GPU	Sehr hohe Generation bis ~258 tok/s
RAG / große Prompts	Qwen3.5-9B Full-GPU	Stärkstes Prompt Processing (~10,5–10,9k tok/s bei 32K)
Sehr langer Kontext	Gemma 4 12B Full-GPU	Beste relative Stabilität bis 262K
Mehrere Nutzer	Gemma 4 12B Server	Bis 1,18–1,41k System-tok/s bei 32 Requests
Zwei Modelle parallel	Qwen3.5-9B + Gemma 4	Bis 1,33k tok/s kombiniert im Windows-Lauf
119B/122B lokal	CPU + großer RAM	Brauchbar; 8–20 tok/s je nach Modell/OS
119B/122B erzwungen Full-GPU	Unified-/System-Memory-Overload	Nicht empfehlenswert; deutlicher Leistungseinbruch

8. Gesamtfazit

Fazit für den Betrieb

Das System eignet sich sehr gut als leistungsstarker lokaler LLM-Arbeitsplatz und als kleiner Inferenzserver. Die RTX 5090 bietet für Modelle bis etwa 27B bzw. rund 17 GiB GGUF-Gewicht sehr hohe Leistung und starke Parallelität. Die 32-GB-VRAM-Grenze ist der maßgebliche Engpass für 70-GiB-Modelle, wird aber durch den Threadripper und ca. 256 GB RAM teilweise kompensiert. Ubuntu und Windows liefern bei der GPU-Einzelinferenz nahezu gleichwertige Ergebnisse; Unterschiede werden stärker bei CPU-Inferenz und bestimmten Serving-Szenarien sichtbar. Thermik und Stromversorgung zeigen in beiden Testreihen ausreichende Reserven für dauerhafte Hochlast.

- Ubuntu: leichte Vorteile bei Single-Stream-GPU und deutlicher Vorteil im 16-/32-Request-TTFT-Serving-Test.
- Windows: klare Vorteile bei CPU-Generation und im getesteten Zwei-Modell-Multi-Tenant-Szenario.
- Long Context: weitgehend plattformneutral; Modellarchitektur beeinflusst die Skalierung deutlich stärker als das Betriebssystem.
- Dauerlast: beide Plattformen stabil, ohne relevanten thermischen Durchsatzverlust.

Anhang A – Kennzahlen auf einen Blick

Kennzahl	Ubuntu	Windows
Laufzeit Referenztest	19,30 h	38,98 h
Bestes GPU tg128	257,7 tok/s (GLM)	253,7 tok/s (GLM)
Bestes PP 32K	10.908 tok/s (Qwen3.5-9B)	10.498 tok/s (Qwen3.5-9B)
Gemma @ 262K	107,8 tok/s	105,8 tok/s
Qwen3.5-9B @ 262K	102,0 tok/s	100,9 tok/s
TTFT C32 Systemdurchsatz	1.409,6 tok/s	1.176,9 tok/s
TTFT C32 P95	857 ms	893 ms
Multi-Tenant C8 kombiniert	954,5 tok/s	1.330,4 tok/s
Max. GPU-Temperatur	72 °C (Server-Stress)	72 °C (Server-Stress)
GPU-Leistung Spitze	608,0 W (Multi-Tenant)	604,6 W (TTFT)
OOM-Test: beständig stabil	mind. 130.012 Tokens	mind. 130.012 Tokens

Anhang B – Abkürzungen und Begriffserklärungen

Die folgende Tabelle erläutert die im Bericht verwendeten technischen Abkürzungen und Fachbegriffe in allgemeinverständlicher Form.

Abkürzung / Begriff	Bedeutung	Erklärung
LLM-Inferenz	Ausführung eines Sprachmodells	Nutzung eines bereits trainierten Modells zur Verarbeitung einer Eingabe und Erzeugung einer Antwort.
16C/32T	16 Cores / 32 Threads	Kurzangabe für 16 Prozessorkerne und 32 gleichzeitig nutzbare Ausführungsthreads des Threadripper-Prozessors.
CUDA	Compute Unified Device Architecture	NVIDIA-Plattform, über die Anwendungen die Rechenleistung einer NVIDIA-GPU für allgemeine Berechnungen und KI nutzen können.
GGUF	Modell-Dateiformat	Dateiformat für lokal ausführbare Sprachmodelle, das insbesondere von llama.cpp verwendet wird.
llama.cpp	LLM-Inferenzsoftware	Software zur effizienten lokalen Ausführung von Sprachmodellen auf CPU und GPU. Unterschiedliche Builds können die Leistung beeinflussen.
llmbench	LLM-Benchmarksoftware	Im Bericht verwendetes Benchmarkprogramm zur automatisierten Messung von Modell-, Kontext-, Server- und Dauerlastleistung.
tok/s	Tokens pro Sekunde	Geschwindigkeitsangabe für Sprachmodelle. Sie zeigt, wie viele Tokens pro Sekunde verarbeitet oder erzeugt werden. Höhere Werte bedeuten mehr Durchsatz.
Token	Verarbeitungseinheit eines Sprachmodells	Text wird von einem LLM in kleinere Einheiten zerlegt. Ein Token kann ein Wort, ein Wortteil, ein Zeichen oder eine ähnliche Texteinheit sein.
tg	Token Generation	Messung der Geschwindigkeit, mit der das Modell neue Ausgabe-Tokens erzeugt.
tg128	Token Generation, 128 Tokens	Generierungsleistung bei einer Ausgabe von 128 neuen Tokens.
tg1024	Token Generation, 1024 Tokens	Generierungsleistung bei einer längeren Ausgabe von 1024 neuen Tokens.
PP	Prompt Processing	Geschwindigkeit, mit der das Modell die Eingabe verarbeitet, bevor die eigentliche Antwortgenerierung beginnt.
PP 32K	Prompt Processing bei 32K Tokens	Prompt-Verarbeitungsleistung bei einer Eingabe von ungefähr 32.000 Tokens.
Prompt	Eingabe an das KI-Modell	Text, Anweisung, Dokumentinhalt oder sonstige Eingabe, die dem Sprachmodell zur Verarbeitung übergeben wird.
Context / Kontext	Verarbeitbarer Textbereich	Gesamter Textbereich, den das Modell innerhalb einer Anfrage berücksichtigen kann, einschließlich Eingabe und vorherigem Gesprächsverlauf.
K	Tausend	Abkürzung für ungefähr tausend Tokens. 32K steht beispielsweise für rund 32.000 Tokens, 262K für rund 262.000 Tokens.
Long Context	Sehr großer Kontext	Test, der untersucht, wie sich Geschwindigkeit und Stabilität des Modells bei sehr langen Eingaben verändern.
KV-Cache	Key-Value Cache	Zwischenspeicher für bereits berechnete Kontextinformationen eines Transformer-Modells. Mit größerem Kontext steigt in der Regel sein Speicherbedarf.
F16 / FP16	16-Bit-Gleitkommazahl	Zahlenformat mit 16 Bit Genauigkeit. Im Referenztest wurde der KV-Cache im F16-Format betrieben.
Batch	Verarbeitungsblock	Menge von Daten bzw. Tokens, die gemeinsam verarbeitet werden. Größere Batches können die Hardware besser auslasten, benötigen jedoch mehr Speicher.
UBatch	Micro-Batch	Interne Unterteilung eines größeren Batches in kleinere Verarbeitungseinheiten, um Speicherverbrauch und Ausführung besser zu steuern.
Flash Attention	Optimiertes Attention-Verfahren	Speicher- und rechenoptimierte Umsetzung eines zentralen Bestandteils moderner Transformer-Sprachmodelle.
Full-GPU	Modell vollständig auf der GPU	Das Modell befindet sich vollständig im VRAM und wird überwiegend bzw. vollständig auf der GPU berechnet. Dies liefert üblicherweise die höchste Leistung.
CPU-only	Ausschließliche CPU-Ausführung	Das Modell wird ohne relevante GPU-Beschleunigung auf dem Hauptprozessor ausgeführt.
Offloading	Auslagerung von Modellteilen	Teile eines Modells werden zwischen GPU und Arbeitsspeicher bzw. CPU verteilt, wenn das Modell nicht vollständig in den VRAM passt.
Unified-/System-Memory-Overload	Ausweichen auf System-RAM bei VRAM-Mangel	Situation, in der ein für den VRAM zu großes Modell teilweise über den normalen Arbeitsspeicher bedient wird. Dies kann die Leistung stark reduzieren.
OOM	Out of Memory	Bedeutet, dass für eine Berechnung nicht genügend Speicher verfügbar ist. Im Bericht bezeichnet der OOM-/Kapazitätstest die überprüfte Kontext- bzw. Speichergrenze.

Abkürzung / Begriff	Bedeutung	Erklärung
Concurrency	Anzahl gleichzeitiger Anfragen	Gibt an, wie viele Requests der Server gleichzeitig bearbeitet. Concurrency 32 bedeutet 32 parallele Anfragen.
C8 / C16 / C32	Concurrency 8 / 16 / 32	Kurzschreibweise für 8, 16 bzw. 32 gleichzeitig bearbeitete Anfragen.
System-tok/s	Gesamtdurchsatz des Servers	Gesamtzahl der pro Sekunde erzeugten Tokens über alle gleichzeitig laufenden Anfragen hinweg.
tok/s je Request	Durchsatz pro Anfrage	Durchschnittliche Tokenrate, die eine einzelne Anfrage bei paralleler Serverlast erhält.
TTFT	Time To First Token	Zeit vom Absenden einer Anfrage bis zum ersten vom Modell ausgegebenen Token. Ein kleinerer Wert bedeutet eine schnellere wahrgenommene Reaktion.
P50	50. Perzentil / Median	50 % der Messwerte liegen auf oder unter diesem Wert. Er beschreibt die typische Latenz einer Anfrage.
P95	95. Perzentil	95 % der Messwerte liegen auf oder unter diesem Wert. Er zeigt, wie hoch die Latenz in ungünstigeren Fällen ausfällt.
ms	Millisekunden	Zeiteinheit. 1.000 ms entsprechen einer Sekunde.
Multi-Tenant	Parallele Nutzung durch mehrere Instanzen	Betriebsszenario, in dem mehrere Nutzer, Anwendungen oder Modelle gleichzeitig Ressourcen desselben Systems verwenden.
Single-Stream	Einzelner Inferenzstrom	Messung mit einer einzelnen Anfrage bzw. ohne hohe Parallelität. Sie zeigt die Leistung eines einzelnen Modelllaufs.
Serving	Bereitstellung eines Modells als Dienst	Serverbetrieb, bei dem ein oder mehrere KI-Modelle für andere Nutzer oder Anwendungen Anfragen entgegennehmen.
RAG	Retrieval-Augmented Generation	Verfahren, bei dem ein Sprachmodell zusätzlich mit gefundenen Dokumenten oder Wissensinhalten versorgt wird, bevor es die Antwort erzeugt.
TPS	Tokens Per Second	Englische Bezeichnung für Tokens pro Sekunde. Beim TPS-Verlust wird die Änderung des Durchsatzes während eines längeren Lasttests bewertet.
Thermal Throttling	Temperaturbedingte Leistungsreduzierung	Automatische Reduzierung von Takt bzw. Leistung zum Schutz vor zu hohen Temperaturen. Im Test wurde kein relevanter thermischer Leistungseinbruch festgestellt.
B bei Modellnamen	Billion = Milliarden Parameter	Bei Modellnamen steht z. B. 9B, 27B oder 119B für ungefähr 9, 27 bzw. 119 Milliarden Modellparameter.
A10B	Active 10 Billion Parameters	Kennzeichnung eines Modells, bei dem trotz höherer Gesamtparameterzahl pro Verarbeitungsschritt nur ungefähr 10 Milliarden Parameter aktiv genutzt werden.
Build	Kompilierungs-/Versionsstand	Konkreter Entwicklungs- oder Kompilierungsstand einer Software. Unterschiedliche Builds können Leistungsunterschiede verursachen.
Performance-Governor	CPU-Leistungsprofil unter Linux	Linux-Einstellung, die die CPU auf hohe und schnelle Leistungsbereitstellung statt auf maximale Energieeinsparung ausrichtet.
Native Ausführung	Direkte Ausführung im Betriebssystem	Die Anwendung läuft direkt unter dem jeweiligen Betriebssystem und nicht innerhalb eines Containers.
Referenzlauf	Standardisierter Haupttestlauf	Benchmarkdurchlauf, dessen Messwerte im Bericht für den direkten Vergleich der beiden Plattformen verwendet werden.