Six of eleven baseline models wrote a correct decoder

Across 18 single-model attempts, nine passed, six failed and three timed out. In separate host trials, all three tested 8K routers delegated a passing decoder to a 256K worker.

Apple M2 Max

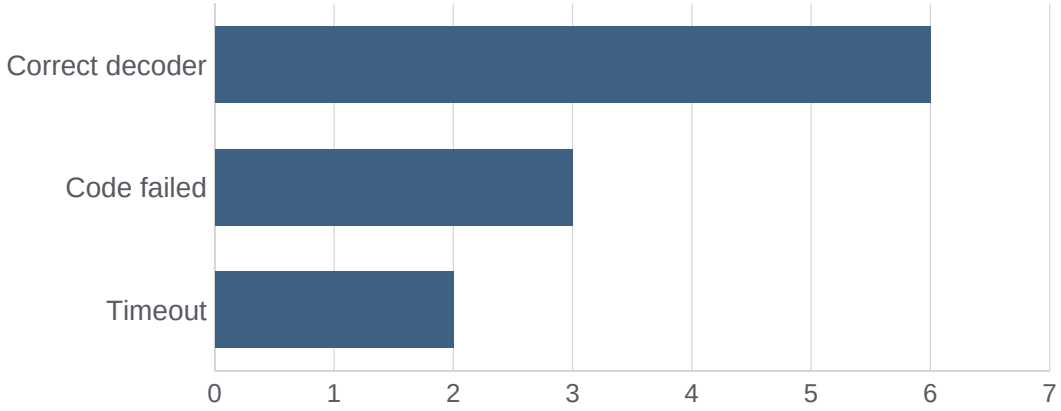
MacBook Pro · Mac14,6

12 CPU cores / 38 GPU cores

64 GB unified memory

macOS 27.0.1 (26A434) · Metal

Baseline outcomes: one trial for each of 11 models



Fastest passing run

Qwen3.6 35B-A3B Q8_0 with MTP: 155.3 s to generate, 180.2 s through verification.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

How to reproduce and read this benchmark

The baseline varies the GGUF configuration while keeping the task, output contract and oracle fixed. One generation per configuration; a second oracle execution uses the same saved code.

Protocol	Setting
Engine	llama.cpp b12649 · Metal · offline · one resident model and one slot
Task / context	hw-sensor-frames-python · 32,768 token context · seed 724
Batches / KV	Batch 512 · microbatch 128 · K and V cache q8_0
Output limit	24,576 generated tokens · 900 seconds · model-specific sampling
Oracle	Run decoder.py on six seeded byte streams · verifier seed 20261018
Pass condition	Program output must match an independent reference on every stream.

What the numbers mean

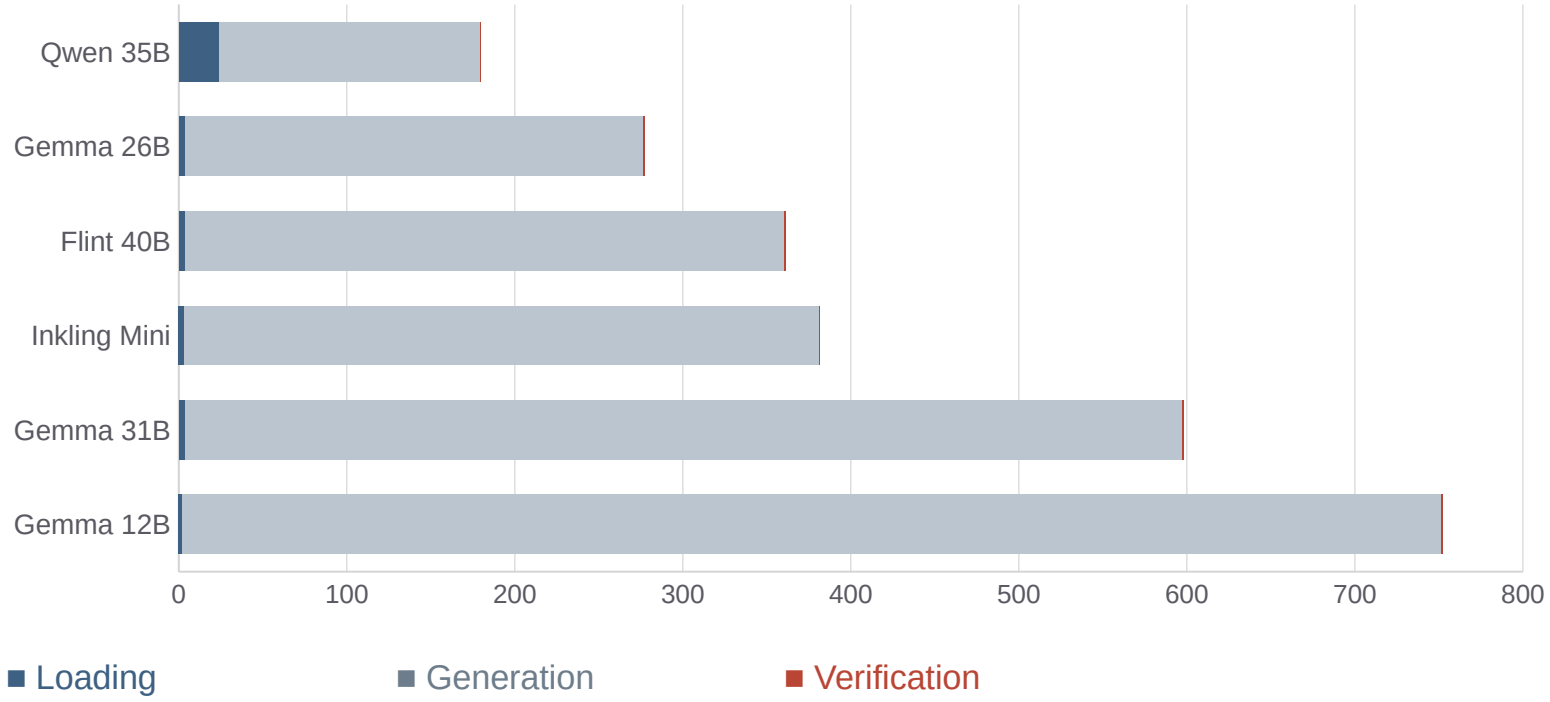
Times and token counts come from completed runs. GPU memory is a dry-run estimate. System load can disqualify a run as a timing reference. Historical 256K tests and the MTP A/B are separate series.

Hardware identity from sysctl and system_profiler. Paired resident-model tests were pending at the snapshot time.

Three minutes from loading to a verified decoder

Qwen3.6 35B-A3B Q8_0 with multi-token prediction (MTP) finished in 180.2 seconds. Generation took 155.3 seconds.

Time by stage for the six passing models, seconds



180.2 seconds

The other five passing runs took 277.6 to 752.7 seconds. Loading Qwen took 24.2 seconds. The next table gives the stage timings.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

Where the time went

One attempt per configuration. Total time includes the brief transitions between stages.

Model	Weights	Load, s	Generation, s	Verification, s	Total, s
Qwen3.6 35B-A3B	Q8_0	24.18	155.3	0.53	180.2
Gemma 4 26B-A4B	Q8_0	4.04	272.6	0.75	277.6
Flint 40B	UD-Q6_K_XL	4.03	356.3	0.78	361.3
Inkling 2 Mini	Q8_0	3.03	377.9	0.47	381.5
Gemma 4 31B	Q8_0	4.04	593.3	0.65	598.2
Gemma 4 12B	Q8_0	2.03	749.5	1.10	752.7

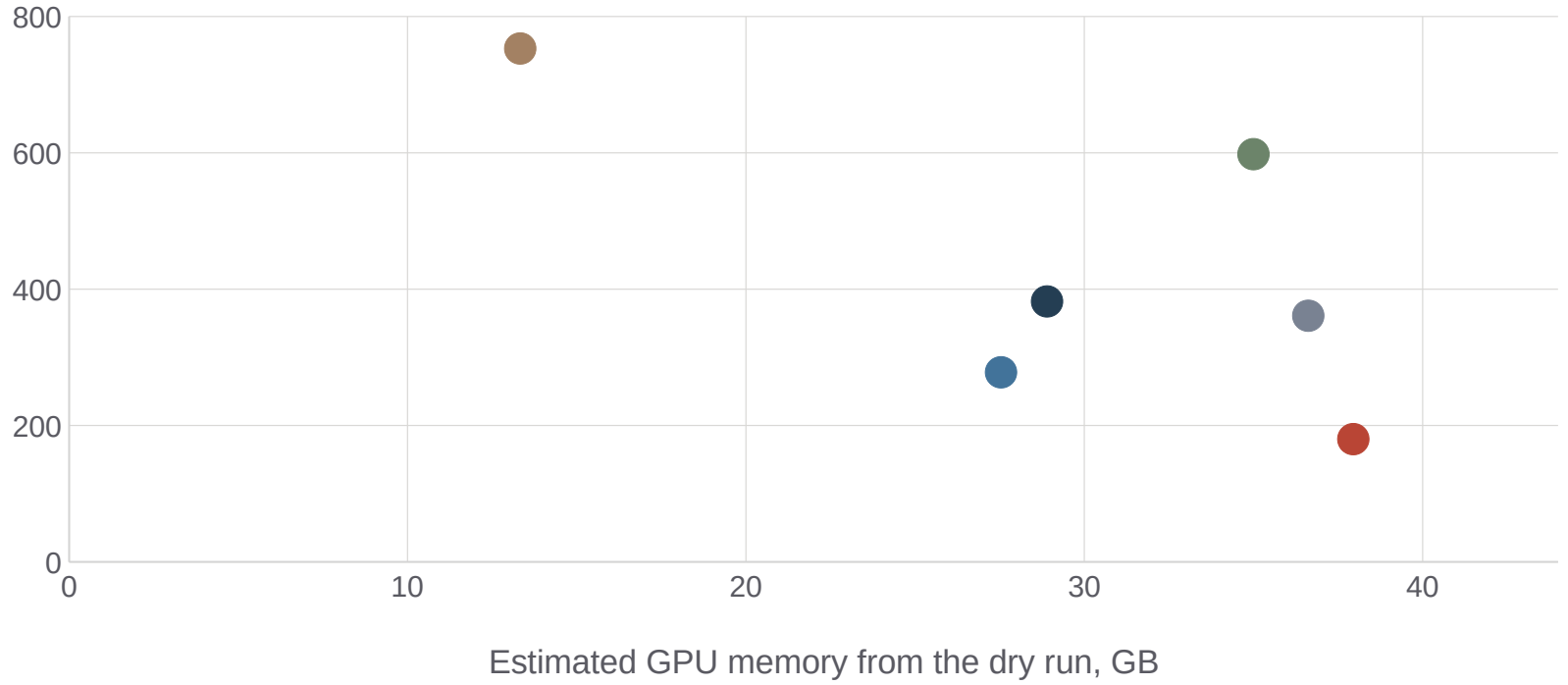
System load exceeded the reference timing threshold during the Qwen3.6 35B and Flint runs. These durations describe the observed runs.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

Memory and time to a verified solution

Each point is a passing decoder. The original dry run measured the target model only. The appendix includes the MTP head or context where available.

Total time through verification, seconds



Model / GB and seconds

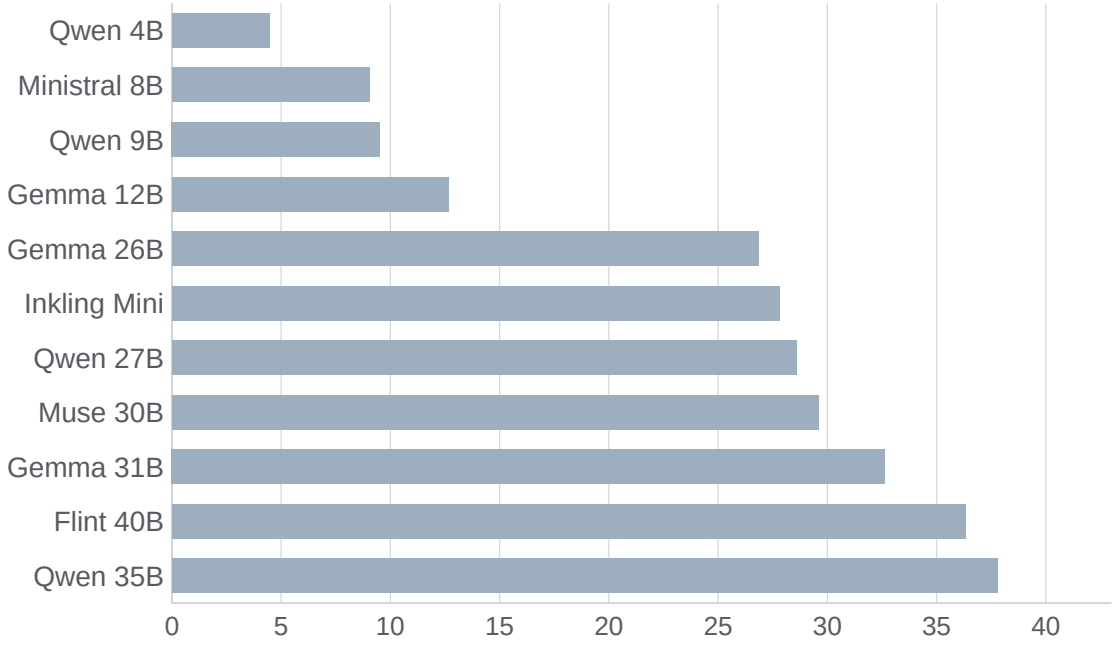
- **Qwen 35B 38.0 / 180**
- Gemma 26B 27.5 / 278
- Flint 40B 36.6 / 361
- Inkling Mini 28.9 / 382
- Gemma 31B 35.0 / 598
- Gemma 12B 13.3 / 753

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

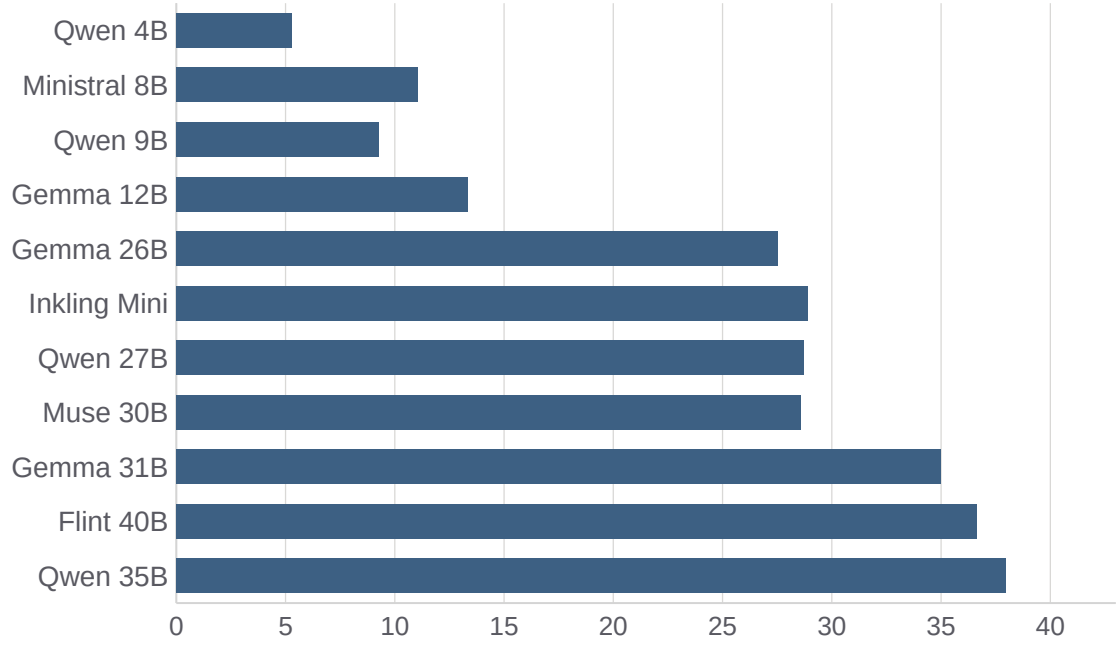
Model weights on disk and memory at runtime

Runtime memory includes weights, context cache and compute buffers. The 32K estimate is for a single server instance.

A. GGUF file size, decimal GB



B. Runtime estimate at 32K, GB



The original 37.95 GB figure is the target-only estimate. The later memory matrix includes the MTP context or head and distinguishes Metal from host memory.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

How token count and throughput affect generation time

For runs that produced a program, throughput ranged from 5.87 to 60.11 tokens per second. Some models finished sooner despite generating more tokens.

Model	Tokens generated	Tokens/s	Generation, s
Qwen3.6 35B-A3B	9,240	60.11	155.3
Flint 40B	15,253	43.07	356.3
Gemma 4 26B-A4B	10,012	36.93	272.6
Qwen3.5 4B	13,491	34.44	392.7
Qwen3.5 9B	1,925	26.37	75.0
Ministral 3 8B Instruct	1,238	23.74	54.0
Inkling 2 Mini	8,364	22.42	377.9
Gemma 4 12B	10,053	13.48	749.5
Gemma 4 31B	3,430	5.87	593.3

Two runs compared

Gemma 4 31B generated 3,430 tokens in 593.3 s. Flint generated 15,253 in 356.3 s. Both counts include reasoning before the code.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

Three programs failed and two runs timed out

Generation was capped at 900 seconds. Neither timed-out run produced a program to verify or a time to solution.

Model · Q8_0	Outcome	Generation, s	Total, s	Independent check
Ministral 3 8B Instruct	Code failed	54.0	57.3	0 output lines for 300 valid frames
Qwen3.5 9B	Code failed	75.0	77.6	GPS field too short to unpack
Qwen3.5 4B	Code failed	392.7	394.4	Tried to concatenate a list and bytes
Qwen3.8 27B	Timeout	900.0	—	6,586 tokens observed, no file
Muse Glimmer 30B	Timeout	900.5	—	6,444 tokens observed, no file

The verifier executed all three failing programs. The timed-out runs recorded partial token counts and neither produced a decoder.py file.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

All eleven configurations and their missing measurements

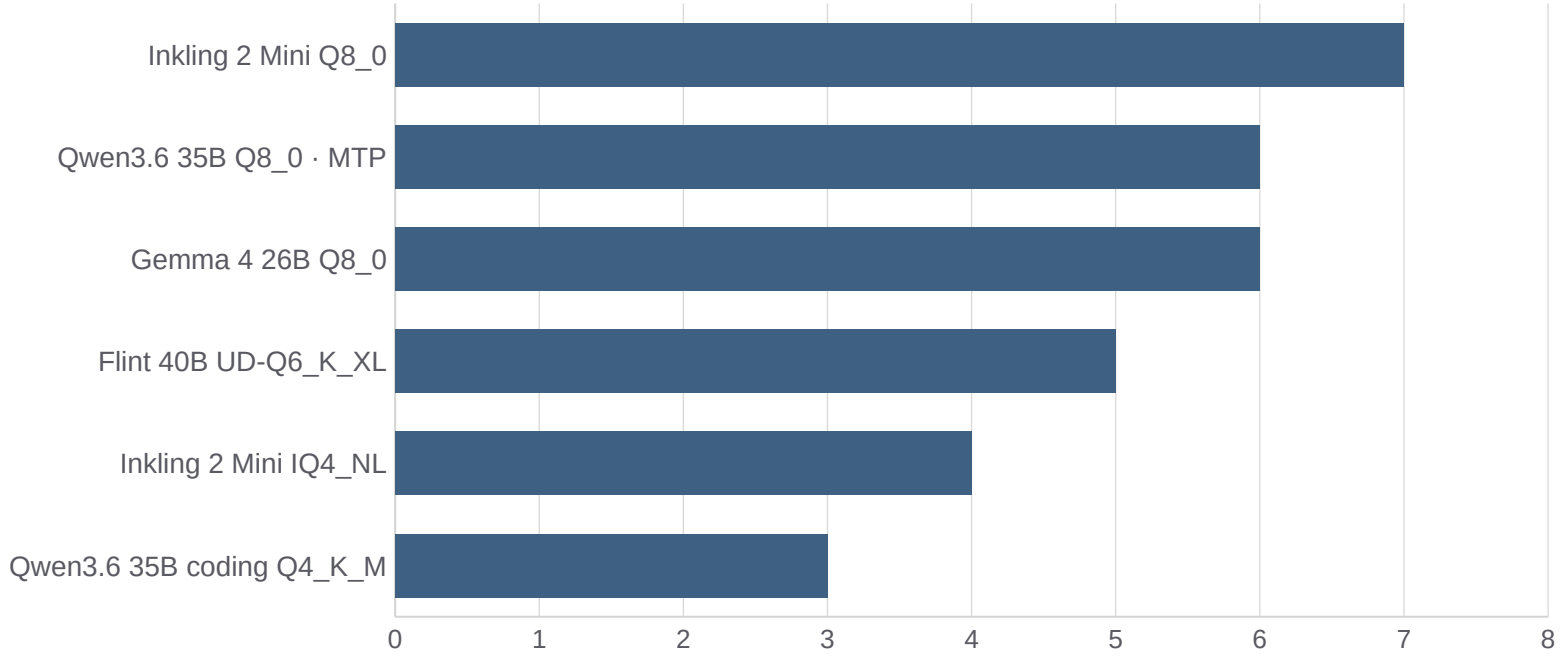
File size is in decimal GB. The GPU column gives the 32K dry-run estimate. A dash marks a missing measurement or stage.

Model	Weights	GGUF GB	GPU GB	Load s	Gen. s	Total s	Tokens/s
Qwen3.6 35B-A3B	Q8_0	37.80	37.95	24.2	155.3	180.2	60.11
Gemma 4 26B-A4B	Q8_0	26.86	27.54	4.0	272.6	277.6	36.93
Flint 40B	UD-Q6_K_XL	36.32	36.62	4.0	356.3	361.3	43.07
Inkling 2 Mini	Q8_0	27.82	28.90	3.0	377.9	381.5	22.42
Gemma 4 31B	Q8_0	32.64	35.00	4.0	593.3	598.2	5.87
Gemma 4 12B	Q8_0	12.67	13.33	2.0	749.5	752.7	13.48
Ministral 3 8B Instruct	Q8_0	9.03	11.05	2.0	54.0	57.3	23.74
Qwen3.5 9B	Q8_0	9.53	9.25	2.0	75.0	77.6	26.37
Qwen3.5 4B	Q8_0	4.48	5.28	1.0	392.7	394.4	34.44
Qwen3.8 27B	Q8_0	28.60	28.72	—	900.0	—	—
Muse Glimmer 30B	Q8_0	29.61	28.58	—	900.5	—	—

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

At 256K, Inkling solved seven of eight tasks

The earlier series tested eight different tasks. Each bar counts tasks solved on the first attempt.



262,144-token context window · microbatch sizes and seeds varied across earlier runs

Observed difference

Inkling Q8 solved one more task than the two runners-up and three more than its IQ4_NL variant. The single 32K task measures a different capability.

Source: memoria-modelli/RAPPORTO.md, per-model results and individual task logs

What remains open for the three tiers

This snapshot covers one attempt at one task. Choosing default models also requires testing how they work together.

Tier	Candidate	Open question at 09:01 CET
Light	Selection open	Routing, when to delegate, and the cost of keeping a worker resident.
Medium	Gemma 4 26B-A4B Q8_0	Cost of pairing with Light and the effect of an MTP head.
Heavy	Qwen3.6 35B-A3B Q8_0	MTP on/off with the same weights, residency alongside Light, and coverage on additional requests.

Qwen finished the 32K decoder first. Inkling solved the most tasks in the 256K series. Both results matter when choosing a tier.

Source: MISURE.tsv, single-model run receipts and the GGUF catalog

Supplement / The same Qwen model with MTP on and off

Qwen3.6 35B-A3B Q8_0 ran with identical weights, prompt, seed 724 and 32K context. Both decoders passed all six streams.

Mode	Generation	Tokens generated	Drafts accepted / proposed	Δ wired at load
MTP on	155.3 s	9,240	5,822 / 6,834	38.70 GB
MTP off	302.2 s	11,454	0 proposals observed	37.87 GB

The MTP-on run finished 146.9 s sooner. Output lengths differed and system load exceeded the reference threshold during that run. A single pair cannot establish a repeatable speedup.

Sources: single-model MTP run receipt and independent A/B audit (heavyq8)

Eighteen 32K runs: nine passed, six failed, three timed out

The eleven baselines use eleven different configurations. Seven additional runs toggle MTP on the same primary weights. A pass requires correct output on all six streams.

Series	Runs	Passed	Failed programs	Timed out
Original single-model baseline	11	6	3	2
Additional MTP variants	7	3	3	1
Combined attempts	18	9	6	3

The three additional binary-stream passes came from Qwen3.6 35B, Gemma 4 26B and Gemma 4 12B. Gemma 4 31B switched from pass to fail on the MTP-on sample.

Sources: 18 single-model receipts and independent artifact replays on 15 saved programs

The same decoder at a 262,144-token window

Six earlier runs covered four families. The historical eight-task score measures breadth; this table isolates the decoder task from that series.

Model	Weights	Outcome	Generation, s	Generated tokens	Microbatch
Qwen3.6 35B-A3B	Q8_0	PASS	186.5	11,038	512
Inkling 2 Mini	Q8_0	PASS	208.1	4,635	128
Gemma 4 26B-A4B	Q8_0	PASS	222.3	9,106	512
Inkling 2 Mini	IQ4_NL	PASS	229.8	4,507	128
Qwen3.6 35B-A3B coding	Q4_K_M	FAIL	271.8	13,434	512
Flint 40B	UD-Q6_K_XL	PASS	277.1	13,351	512

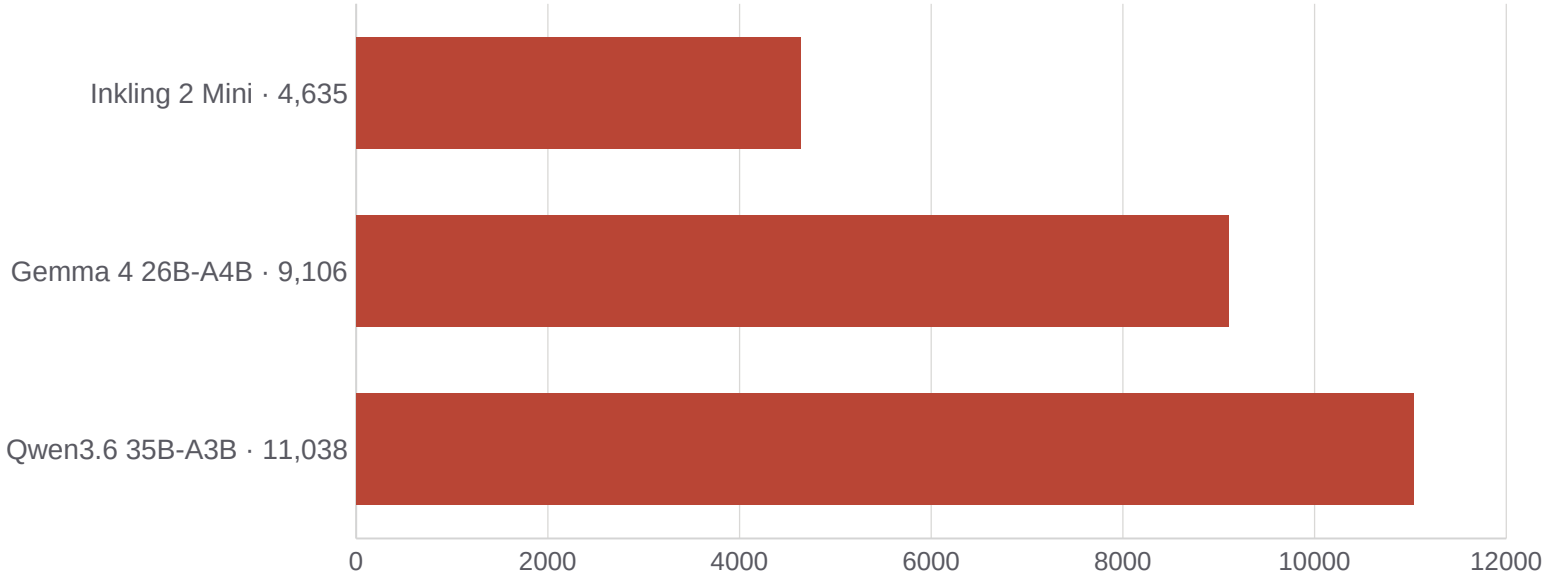
Five of six historical programs passed. Earlier sampling and microbatches varied. A difference from the 32K results cannot be attributed to the context window alone.

Sources: historical 256K JSONL entries and Flint single-task evidence

At 256K, Inkling used 4,635 tokens for a passing decoder

Among the three passing Q8_0 runs, Inkling 2 Mini generated the fewest tokens. Each run received the same decoder task and was scored on the same six streams.

Generated tokens, including reasoning and code



Other precisions

Inkling 2 Mini IQ4_NL also passed, using 4,507 tokens. Flint 40B UD-Q6_K_XL passed using 13,351. Their weight formats differ from the three Q8_0 runs.

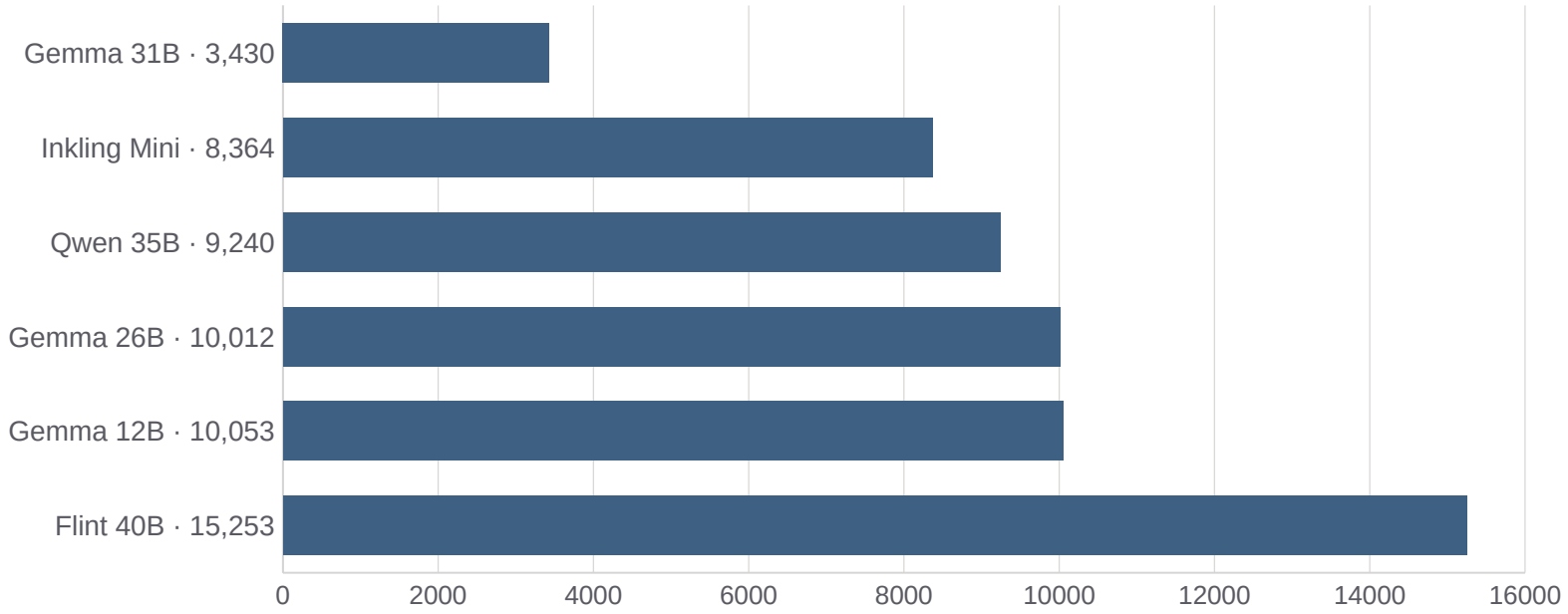
Historical sampling settings and microbatch sizes varied. These counts describe the observed generations.

Source: historical decoder JSONL entries at 262,144 tokens. One generation per configuration.

At 32K, Gemma 4 31B used 3,430 tokens

All six passing baseline runs are shown. Inkling 2 Mini Q8_0 used 8,364 tokens. The ordering on this task changed across contexts and weight configurations.

Generated tokens per passing run



Count and speed

Gemma 4 31B generated the shortest passing response but took 593.3 seconds. Qwen3.6 35B-A3B generated 9,240 tokens in 155.3 seconds.

Flint uses UD-Q6_K_XL. All five other entries use Q8_0. A single generation does not establish a general token-efficiency ranking.

Seven MTP variants extend the eleven-model baseline

Each A/B retains the primary GGUF, decoder task, seed 724 and 32K context. Only configurations with a supported MTP method enter this comparison.

Model · Q8_0	MTP base → variant	Oracle base → variant	Generation base → variant, s	Drafts accepted / proposed[1]
Gemma 4 12B	off → on	pass → pass	749.5 → 381.8	6,627 / 8,340
Gemma 4 26B-A4B	off → on	pass → pass	272.6 → 199.6	5,800 / 7,054
Gemma 4 31B	off → on	pass → fail	593.3 → 405.2	3,109 / 3,884
Qwen3.5 4B	off → on	fail → fail	392.7 → 293.7	9,315 / 12,376
Qwen3.5 9B	off → on	fail → fail	75.0 → 52.1	1,205 / 1,438
Qwen3.6 35B-A3B	on → off	pass → pass	155.3 → 302.2	5,822 / 6,834
Qwen3.8 27B	off → on	timeout → timeout	900.0 → 900.3	—

[1] Counters refer to the MTP-on run. A timeout may retain runtime evidence of active speculation without final acceptance counters. An unsuccessful run has no time to a correct solution.

Sources: recorded MTP manifests, runtime counters, code hashes and independent oracle replays

How much memory changed with MTP

Absolute change in machine-wide wired memory after model loading relative to the immediately preceding sample. Decimal GB. These are measured system differences.

Model · Q8_0	Wired Δ, off	Wired Δ, on	Difference on − off	Reference timing: off / on
Gemma 4 12B	14.02	14.69	0.68	yes / yes
Gemma 4 26B-A4B	28.40	29.12	0.71	yes / yes
Gemma 4 31B	35.67	36.64	0.97	yes / no
Qwen3.5 4B	5.93	7.32	1.40	no / no
Qwen3.5 9B	10.01	11.99	1.98	no / no
Qwen3.6 35B-A3B	37.87	38.70	0.83	no / no
Qwen3.8 27B	—	—	—	— / —

A smaller time with unchanged or failed correctness does not establish a universal speedup. Output trajectories and lengths can differ even on the same task.

Sources: paired memory snapshots and runtime logs; single observation per mode

Four baseline configurations lack a supported A/B

The test fixes the main weights and llama.cpp b12649. Different speculative decoding schemes require separate compatibility checks.

Model	Primary weights	Eligibility in this protocol
Inkling 2 Mini	Q8_0	DSpark drafter; a compatible standard MTP head for b12649 was not verified.
Muse Glimmer 30B	Q8_0	DFlash drafter; no standard MTP head was qualified for these weights.
Flint 40B	UD-Q6_K_XL	No compatible standard MTP head was verified for the measured weight file.
Ministral 3 8B Instruct	Q8_0	No compatible standard MTP head was verified for the measured weight file.

All four retain their baseline results. The absence of a compatible A/B is recorded separately from a program that fails the verifier.

Source: MTP head inventory, upstream model repositories and runtime admission checks

The production router and the final answer

The Rust host chooses among four real tools. The Light router uses 8K, chat reloads at 32K, and a 256K worker is eligible only when the configuration fits.

Router / weights	Request	Action	Observed artifact	Router s	Request s
Ministral 3 8B Instruct Q8_0	Text extraction	chat	fail	2.2	4.8
Ministral 3 8B Instruct Q4_K_M	Python decoder	develop	pass	5.9	251.2
Qwen3.5 4B Q8_0	Text extraction	chat	pass	2.9	7.6
Qwen3.5 4B Q8_0	Python decoder	develop	pass	9.1	163.2
Qwen3.5 9B Q8_0	Text extraction	chat	pass	8.8	23.7
Qwen3.5 9B Q4_K_M	Python decoder	develop	pass	15.5	220.9

All six routing decisions matched the expected action. Two of three extractions returned exactly “ambra”. All three delegated decoders passed six streams; their request times ranged from 163.2 to 251.2 seconds.

Sources: production host payload, llama.cpp argv, model responses and independent artifact replays

Router context and observed model residency

The quick decision reserves 320 output tokens. The thinking call reserves 600. The prompt and reserve must both fit; the minimum measured configuration is distinct from a theoretical lower bound.

Model / weights	Request	Router ctx	Maximum required[1]	Chat / worker ctx[2]	Both observed
Ministral 3 8B Instruct Q8_0	Extraction	8,192	1,573	32,768	No
Ministral 3 8B Instruct Q4_K_M	Decoder	8,192	2,486	262,144	Yes
Qwen3.5 4B Q8_0	Extraction	8,192	1,781	32,768	No
Qwen3.5 4B Q8_0	Decoder	8,192	2,657	262,144	Yes
Qwen3.5 9B Q8_0	Extraction	8,192	1,781	32,768	No
Qwen3.5 9B Q4_K_M	Decoder	8,192	2,657	262,144	Yes

[1] Maximum of prompt tokens plus reserved output across the two host payloads. [2] Context of the final serving process when observed. A correct routing decision with no worker available produces no verified handoff.

Sources: rendered prompt token counts, router limits and listener-verified process samples

A matched 256K worker finished alone in 209.4 seconds

Gemma 4 26B-A4B Q8_0 with MTP passed every stream alone and with each Light router. All three routers used an 8,192-token context.

Router / weights	Routing, s	Worker, s	Request, s	Difference vs solo	Oracle
None · solo control	—	209.4	209.4	—	Pass
Qwen3.5 4B Q8_0	9.1	154.0	163.2	-46.2 s	Pass
Qwen3.5 9B Q4_K_M	15.5	205.4	220.9	+11.5 s	Pass
Ministral 3 8B Instruct Q4_K_M	5.9	245.3	251.2	+41.8 s	Pass

The solo worker produced the same response and decoder at 32K (199.6 s) and 256K (209.4 s). Each paired run added a different router assignment to the original task. The time differences include changes to generation as well as the cost of routing.

Sources: matched solo receipt, production routing receipts and independent replays of all four decoders

The same router helped one Heavy worker finish in time

Qwen3.5 4B Q8_0 with MTP routed the decoder request at 8K. Each worker ran at 256K on the 64 GB M2 Max. Seed 724; one generation per condition.

Worker / quantization	Worker alone	With 4B router	What the six streams showed
Gemma 4 26B-A4B Q8_0 · MTP	209.40 s · pass	163.16 s · pass	Both decoders passed all six streams.
Qwen3.6 35B-A3B coding Q4_K_M · MTP	197.60 s · fail	168.36 s · pass	The solo decoder scaled acceleration by 1/1,000 on three streams.
Inkling 2 Mini UD-Q8_K_XL	711.40 s · pass	900 s cap · incomplete	The solo decoder passed all six. The paired run produced no file by the cutoff.

All three handoffs used the same router model. The worker saw the original request plus a short assignment generated by the router. With one trial per condition, this comparison cannot isolate the effect of routing on correctness or speed.

What the two new Heavy runs generated

Both paired runs first chose the developer tool. The same external verifier was replayed on fresh copies of all three delivered Python files.

Worker / condition	Route, s	Worker tokens	Worker tok/s	System wired[1]	Result
Qwen3.6 Q4_K_M · pair	9.24	11,191	71.18	39.99 GB	6/6
Qwen3.6 Q4_K_M · solo	—	10,948	55.85	33.44 GB	3/6
Inkling UD-Q8_K_XL · pair	9.42	≥8,669[2]	10.10[2]	36.70 GB	No file
Inkling UD-Q8_K_XL · solo	—	7,584	10.94	29.12 GB	6/6

[1] Maximum sampled system-wide wired memory. Pairs were sampled throughout; singles used three snapshots. [2] Last periodic server sample. The pair reached its shared 900-second request limit. Its cold end-to-end wall time was 909.18 seconds.

Why the Heavy tests used these GGUF variants

The 4B router and its MTP head reserve an estimated 5.746 GB on Metal. Both processes had to fit the 41.2 GB Metal budget before the runner started.

Candidate at 256K	Metal estimate with 4B router	Selection and runtime evidence
Qwen3.6 35B-A3B coding Q4_K_M	32.024 GB	Q8_0 with integrated MTP exceeds budget even alone (42.109 GB). The inspected Q6 candidate lacks an MTP head. Q4_K_M ran with integrated MTP.
Inkling 2 Mini UD-Q8_K_XL	27.719 GB	The standard Q8_0 pair needs 44.569 GB. The complete UD-Q8_K_XL file fits; b12649 assigned 10.463 GB of weights to Metal and 12.392 GB to the host backend. No standard MTP head was available.

The allocations share Apple unified memory. Metal estimates, host buffers and observed system wired memory describe different quantities. The Inkling pair reached a 1-minute load average of 22.85. Its throughput and timeout apply to this run.

What can be concluded about the three tiers?

Light, Medium and Heavy are product tiers. The measurements support candidate selection on these specific tasks; they do not establish general performance.

Tier	Supported by these data	Evidence still needed
Light	Qwen3.5 4B Q8_0 selected chat, returned the exact word, and delegated the decoder to a resident Gemma worker.	Broader unseen requests and repeat runs to quantify routing reliability.
Medium	Gemma 4 26B-A4B Q8_0 produced a verified decoder and has a measured MTP A/B.	More programs, longer tasks, thermal variation and realistic native-app sessions.
Heavy	Qwen3.6 35B-A3B Q8_0 was fastest on the original passing decoder runs. Inkling 2 Mini Q8_0 led the historical eight-task series.	Repeat the same configurations across a broader task set and test extended windows separately.

Each model configuration was generated once per condition. Replaying an existing decoder verifies the program again and does not provide a second model sample.

Sources: single-model decoder runs, MTP A/B pairs and the Light-window diagnostics

A dynamic budget for keeping one or two models resident

Deployment proposal. Check the actual machine before each load. Reserve capacity for macOS, existing applications and cache growth; refresh the estimate before admitting a second process.

Decision	Proposed rule
Precision	Test Q8 or the highest precision that fits. Prefer MTP where supported, after checking the head, runtime acceptance and task correctness.
One resident	Compare the full Metal estimate (target plus head) with the lower of the configured GPU limit and the current safe headroom. Check CPU-side memory separately.
Two residents	Add both simultaneous Metal estimates at their real windows. Check host demand, load both, verify distinct listening processes and measure system pressure before continuing.
Pressure / swap	Use the pre-load swap level as baseline. If swap grows materially or memory pressure rises, release the idle model and rerun sequentially.

Giuseppe reported 18–20 GB of swap while using Inkling. This is a user observation of system-wide swap. It cannot be attributed to Inkling alone without synchronized before/after samples.

Sources: no-allocation memory estimates and observed system-wide telemetry; this policy is a proposal

Capacity planning for five Mac memory sizes

Illustrative initial ceilings use roughly 65% of nominal RAM. The measured 64 GB Mac used a 41.2 GB Metal admission budget. Actual pressure may require a lower limit.

Mac RAM	Start cap[1]	Light / 8K routing or 32K chat	Medium / Heavy at 256K	Residency proposal
16 GB	10.4 GB	Qwen3.5 4B Q8_0 MTP	No verified 256K Medium/Heavy candidate under this cap.	One Light.
32 GB	20.8 GB	Qwen3.5 4B Q8_0 MTP	Gemma 4 12B Q8_0 MTP is a 16.93 GB fit candidate; 256K quality unverified.	Sequential; 5.75 + 16.93 GB exceeds cap.
64 GB[2]	41.2 GB	Qwen3.5 4B Q8_0 MTP	Gemma 4 26B Q8_0 MTP at 32.67 GB. Qwen3.6 35B Q8_0 OFF at 39.91 GB.	Verified pairs include 4B + Gemma26 Q8 and 4B + Qwen35 Q4_K_M.
128 GB	83.2 GB	Qwen3.5 4B BF16 MTP: 9.69 GB (8K).	Gemma26 BF16 MTP: 56.32 GB. Qwen35 Q8 MTP: 42.11 GB.	4B + Gemma26 BF16: 66.01 GB. Verify on the target Mac.
192 GB	124.8 GB	Qwen3.5 9B BF16 MTP: 17.69 GB (8K).	Gemma26 BF16 MTP: 56.32 GB. Qwen35 BF16 OFF: 71.91 GB (no verified head).	9B + Qwen35 BF16: 89.60 GB. Benchmark against Q8 MTP.

[1] Decimal GB, initial policy examples only. [2] MacBook Pro M2 Max, Mac14,6. BF16 fits are allocator estimates from sparse metadata. Generation and pressure have not been tested on the 128/192 GB systems.

Source: target-plus-head dry runs at router 8K, chat 32K and worker 256K. Only the 64 GB machine was tested here.

At 64 GB, the MTP head changes which combinations fit

A 256K worker and an 8K router share a 41.2 GB Metal ceiling. Separate CPU-side buffers and system pressure still need monitoring.

Configuration	MTP	Metal estimate	64 GB Mac admission
Qwen3.6 35B Q8_0 · 256K	off / on	39.908 / 42.109 GB	OFF fits alone; ON exceeds the current budget.
Qwen3.8 27B Q8_0 · 256K	off / on	37.709 / 42.086 GB	ON exceeds the current budget.
Gemma 4 26B Q8_0 · 256K	on	32.671 GB	Candidate worker.
Qwen3.5 4B Q8_0 · 8K + Gemma26	on / on	38.417 GB	Candidate pair. Validate routing and output separately.
Qwen3.5 9B Q8_0 · 8K + Gemma26	on / on	42.918 GB	Q8 pair exceeds cap; Q4_K_M router pair estimates 39.529 GB.
Ministral 3 8B Q8_0 · 8K + Gemma26	off / on	41.796 GB	Q8 pair exceeds cap; Q4_K_M router estimates 38.234 GB.

Earlier successful 256K runs remain historical measurements. A successful earlier run does not override a later admission budget based on complete memory accounting.

Source: complete b12649 Metal estimates including the MTP target and head; paired process receipts for actual residency

Reading the complete memory matrix

The catalog has 243 assets: 242 primary candidates and one auxiliary DSpark drafter. Each cell gives Metal GB / host GB (decimal). The CSV preserves exact bytes.

Label	Interpretation
file	The engine evaluated a complete GGUF. This is a dry-run allocation estimate, not observed resident memory during generation.
header	The engine evaluated a metadata-only sparse descriptor. Inference was not performed on these placeholders.
C / H / N/A	C exceeds the model's native context. H means no verified MTP head. N/A includes the auxiliary drafter, which cannot serve as the primary model.
E / P / —	E identifies a failed or unavailable estimate. P is pending in this snapshot. A dash means no recorded cell. See the CSV for the reason.
RAM vs. output	No allocation estimate counts as a model performance trial. Only the specified routing, chat and worker windows were used for the new production-path diagnostic.

estimated: 3,036 · not applicable: 1,824. Total: 4,860 cells.

Source: RAM allocator matrix and per-cell receipts, llama.cpp b12649 in no-allocation mode

Memory matrix / Qwen3.5 9B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	16.1/2.1	16.3/2.1	16.7/2.1	17.4/2.1	18.1/2.1	18.8/2.1	20.2/2.2	21.7/2.2	C	C
BF16	on	header	17.7/3.1	17.9/3.2	18.4/3.2	19.4/3.2	20.4/3.2	21.4/3.3	23.4/3.3	25.4/3.4	C	C
IQ4_NL	off	header	5.1/0.6	5.2/0.6	5.6/0.6	6.3/0.6	7.0/0.6	7.7/0.7	9.2/0.7	10.6/0.7	C	C
IQ4_NL	on	header	6.6/1.7	6.8/1.7	7.3/1.7	8.3/1.7	9.3/1.8	10.3/1.8	12.3/1.9	14.3/1.9	C	C
IQ4_XS	off	header	4.9/0.6	5.0/0.6	5.4/0.6	6.1/0.6	6.8/0.6	7.5/0.7	9.0/0.7	10.4/0.7	C	C
IQ4_XS	on	header	6.4/1.7	6.6/1.7	7.1/1.7	8.1/1.7	9.1/1.8	10.1/1.8	12.1/1.9	14.1/1.9	C	C
Q3_K_M	off	header	4.5/0.5	4.7/0.5	5.0/0.5	5.7/0.5	6.5/0.5	7.2/0.5	8.6/0.6	10.0/0.6	C	C
Q3_K_M	on	header	6.0/1.6	6.3/1.6	6.8/1.6	7.8/1.6	8.8/1.6	9.7/1.7	11.7/1.7	13.7/1.8	C	C
Q3_K_S	off	header	4.1/0.5	4.3/0.5	4.7/0.5	5.4/0.5	6.1/0.5	6.8/0.5	8.2/0.6	9.7/0.6	C	C
Q3_K_S	on	header	5.7/1.6	5.9/1.6	6.4/1.6	7.4/1.6	8.4/1.6	9.4/1.7	11.4/1.7	13.4/1.8	C	C
Q4_0	off	header	5.1/0.6	5.2/0.6	5.6/0.6	6.3/0.6	7.0/0.6	7.7/0.7	9.2/0.7	10.6/0.7	C	C
Q4_0	on	header	6.6/1.7	6.9/1.7	7.4/1.7	8.3/1.7	9.3/1.8	10.3/1.8	12.3/1.9	14.3/1.9	C	C
Q4_1	off	header	5.5/0.7	5.6/0.7	6.0/0.7	6.7/0.7	7.4/0.7	8.1/0.7	9.6/0.8	11.0/0.8	C	C
Q4_1	on	header	7.0/1.7	7.2/1.8	7.7/1.8	8.7/1.8	9.7/1.8	10.7/1.9	12.7/1.9	14.7/2.0	C	C
Q4_K_M	off	header	5.4/0.6	5.5/0.6	5.9/0.6	6.6/0.6	7.3/0.6	8.0/0.7	9.5/0.7	10.9/0.7	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.5 9B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_K_M	off	file	5.3/0.6	5.5/0.6	5.8/0.6	6.6/0.6	7.3/0.6	8.0/0.7	9.4/0.7	10.8/0.7	C	C
Q4_K_M	on	header	6.9/1.7	7.2/1.7	7.7/1.7	8.6/1.7	9.6/1.8	10.6/1.8	12.6/1.9	14.6/1.9	C	C
Q4_K_M	on	file	6.9/1.7	7.1/1.7	7.6/1.7	8.6/1.7	9.6/1.8	10.6/1.8	12.5/1.9	14.5/1.9	C	C
Q4_K_S	off	header	5.1/0.6	5.3/0.6	5.6/0.6	6.3/0.6	7.0/0.6	7.8/0.7	9.2/0.7	10.6/0.7	C	C
Q4_K_S	on	header	6.6/1.7	6.9/1.7	7.4/1.7	8.4/1.7	9.3/1.8	10.3/1.8	12.3/1.9	14.3/1.9	C	C
Q5_K_M	off	header	6.1/0.7	6.3/0.7	6.7/0.7	7.4/0.7	8.1/0.8	8.8/0.8	10.2/0.8	11.7/0.9	C	C
Q5_K_M	on	header	7.7/1.8	7.9/1.8	8.4/1.8	9.4/1.9	10.4/1.9	11.4/1.9	13.4/2.0	15.4/2.1	C	C
Q5_K_S	off	header	5.9/0.7	6.1/0.7	6.5/0.7	7.2/0.7	7.9/0.8	8.6/0.8	10.0/0.8	11.4/0.9	C	C
Q5_K_S	on	header	7.5/1.8	7.7/1.8	8.2/1.8	9.2/1.9	10.2/1.9	11.2/1.9	13.2/2.0	15.1/2.1	C	C
Q6_K	off	header	6.9/0.9	7.1/0.9	7.4/0.9	8.1/0.9	8.8/0.9	9.6/0.9	11.0/1.0	12.4/1.0	C	C
Q6_K	on	header	8.4/1.9	8.7/2.0	9.2/2.0	10.2/2.0	11.1/2.0	12.1/2.1	14.1/2.1	16.1/2.2	C	C
Q8_0	off	header	8.7/1.1	8.9/1.1	9.2/1.1	10.0/1.1	10.7/1.1	11.4/1.2	12.8/1.2	14.2/1.2	C	C
Q8_0	off	file	8.7/1.1	8.9/1.1	9.2/1.1	10.0/1.1	10.7/1.1	11.4/1.2	12.8/1.2	14.2/1.2	C	C
Q8_0	on	header	10.2/2.2	10.5/2.2	11.0/2.2	12.0/2.3	13.0/2.3	14.0/2.3	15.9/2.4	17.9/2.5	C	C
Q8_0	on	file	10.2/2.2	10.5/2.2	11.0/2.2	12.0/2.3	13.0/2.3	14.0/2.3	15.9/2.4	17.9/2.5	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.5 9B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ2_M	off	header	3.5/0.5	3.6/0.5	4.0/0.5	4.7/0.5	5.4/0.5	6.1/0.5	7.6/0.6	9.0/0.6	C	C
UD-IQ2_M	on	header	5.0/1.6	5.3/1.6	5.8/1.6	6.7/1.6	7.7/1.6	8.7/1.7	10.7/1.7	12.7/1.8	C	C
UD-IQ2_XXS	off	header	3.0/0.5	3.2/0.5	3.5/0.5	4.3/0.5	5.0/0.5	5.7/0.5	7.1/0.6	8.5/0.6	C	C
UD-IQ2_XXS	on	header	4.6/1.6	4.8/1.6	5.3/1.6	6.3/1.6	7.3/1.6	8.3/1.7	10.2/1.7	12.2/1.8	C	C
UD-IQ3_XXS	off	header	3.8/0.5	4.0/0.5	4.4/0.5	5.1/0.5	5.8/0.5	6.5/0.5	7.9/0.6	9.4/0.6	C	C
UD-IQ3_XXS	on	header	5.4/1.6	5.6/1.6	6.1/1.6	7.1/1.6	8.1/1.6	9.1/1.7	11.1/1.7	13.1/1.8	C	C
UD-Q2_K_XL	off	header	3.8/0.6	4.0/0.6	4.3/0.6	5.1/0.6	5.8/0.6	6.5/0.7	7.9/0.7	9.3/0.7	C	C
UD-Q2_K_XL	on	header	5.4/1.7	5.6/1.7	6.1/1.7	7.1/1.7	8.1/1.8	9.1/1.8	11.0/1.9	13.0/1.9	C	C
UD-Q3_K_XL	off	header	4.7/0.6	4.9/0.6	5.3/0.6	6.0/0.6	6.7/0.6	7.4/0.7	8.8/0.7	10.3/0.7	C	C
UD-Q3_K_XL	on	header	6.3/1.7	6.5/1.7	7.0/1.7	8.0/1.7	9.0/1.8	10.0/1.8	12.0/1.9	14.0/1.9	C	C
UD-Q4_K_XL	off	header	5.7/0.6	5.8/0.6	6.2/0.6	6.9/0.6	7.6/0.6	8.3/0.7	9.8/0.7	11.2/0.7	C	C
UD-Q4_K_XL	on	header	7.2/1.7	7.4/1.7	7.9/1.7	8.9/1.7	9.9/1.8	10.9/1.8	12.9/1.9	14.9/1.9	C	C
UD-Q5_K_XL	off	header	6.3/0.7	6.5/0.7	6.8/0.7	7.5/0.7	8.3/0.8	9.0/0.8	10.4/0.8	11.8/0.9	C	C
UD-Q5_K_XL	on	header	7.8/1.8	8.1/1.8	8.6/1.8	9.6/1.9	10.6/1.9	11.6/1.9	13.5/2.0	15.5/2.1	C	C
UD-Q6_K_XL	off	header	7.9/1.1	8.1/1.1	8.5/1.1	9.2/1.1	9.9/1.1	10.6/1.2	12.0/1.2	13.5/1.2	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.5 9B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q6_K_XL	on	header	9.5/2.2	9.7/2.2	10.2/2.2	11.2/2.3	12.2/2.3	13.2/2.3	15.2/2.4	17.1/2.5	C	C
UD-Q8_K_XL	off	header	11.2/2.1	11.4/2.1	11.7/2.1	12.4/2.1	13.2/2.1	13.9/2.1	15.3/2.2	16.7/2.2	C	C
UD-Q8_K_XL	on	header	12.7/3.1	13.0/3.2	13.5/3.2	14.5/3.2	15.5/3.2	16.5/3.3	18.4/3.3	20.4/3.4	C	C

Memory matrix / Qwen3.5 4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	8.7/1.3	8.9/1.3	9.2/1.3	9.9/1.3	10.6/1.3	11.4/1.3	12.8/1.4	14.2/1.4	C	C
BF16	on	header	9.7/2.0	9.9/2.0	10.4/2.0	11.4/2.0	12.4/2.1	13.4/2.1	15.4/2.2	17.4/2.2	C	C
IQ4_NL	off	header	2.8/0.5	3.0/0.5	3.4/0.5	4.1/0.6	4.8/0.6	5.5/0.6	6.9/0.6	8.4/0.7	C	C
IQ4_NL	on	header	3.8/1.2	4.1/1.2	4.6/1.2	5.6/1.3	6.6/1.3	7.6/1.3	9.5/1.4	11.5/1.5	C	C
IQ4_XS	off	header	2.7/0.5	2.9/0.5	3.3/0.5	4.0/0.6	4.7/0.6	5.4/0.6	6.8/0.6	8.3/0.7	C	C
IQ4_XS	on	header	3.7/1.2	4.0/1.2	4.5/1.2	5.5/1.3	6.5/1.3	7.5/1.3	9.4/1.4	11.4/1.5	C	C
Q3_K_M	off	header	2.5/0.5	2.7/0.5	3.1/0.5	3.8/0.6	4.5/0.6	5.2/0.6	6.6/0.6	8.1/0.7	C	C
Q3_K_M	on	header	3.6/1.2	3.8/1.2	4.3/1.2	5.3/1.3	6.3/1.3	7.3/1.3	9.2/1.4	11.2/1.5	C	C
Q3_K_S	off	header	2.4/0.5	2.5/0.5	2.9/0.5	3.6/0.6	4.3/0.6	5.0/0.6	6.5/0.6	7.9/0.7	C	C
Q3_K_S	on	header	3.4/1.2	3.6/1.2	4.1/1.2	5.1/1.3	6.1/1.3	7.1/1.3	9.1/1.4	11.0/1.5	C	C
Q4_0	off	header	2.8/0.5	3.0/0.5	3.4/0.5	4.1/0.6	4.8/0.6	5.5/0.6	6.9/0.6	8.4/0.7	C	C
Q4_0	on	header	3.8/1.2	4.1/1.2	4.6/1.2	5.6/1.3	6.6/1.3	7.6/1.3	9.5/1.4	11.5/1.5	C	C
Q4_1	off	header	3.0/0.5	3.2/0.5	3.6/0.5	4.3/0.6	5.0/0.6	5.7/0.6	7.1/0.6	8.6/0.7	C	C
Q4_1	on	header	4.0/1.2	4.3/1.2	4.8/1.2	5.8/1.3	6.8/1.3	7.8/1.3	9.7/1.4	11.7/1.5	C	C
Q4_K_M	off	file	3.3/0.4	3.5/0.4	3.8/0.4	4.5/0.4	5.2/0.4	6.0/0.4	7.4/0.5	8.8/0.5	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.5 4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_K_M	on	file	4.3/1.1	4.5/1.1	5.0/1.1	6.0/1.1	7.0/1.1	8.0/1.2	10.0/1.3	12.0/1.3	C	C
Q4_K_S	off	header	2.8/0.5	3.0/0.5	3.4/0.5	4.1/0.6	4.8/0.6	5.5/0.6	6.9/0.6	8.4/0.7	C	C
Q4_K_S	on	header	3.9/1.2	4.1/1.2	4.6/1.2	5.6/1.3	6.6/1.3	7.6/1.3	9.5/1.4	11.5/1.5	C	C
Q5_K_M	off	header	3.4/0.5	3.6/0.5	3.9/0.5	4.6/0.6	5.4/0.6	6.1/0.6	7.5/0.6	8.9/0.7	C	C
Q5_K_M	on	header	4.4/1.2	4.7/1.2	5.1/1.2	6.1/1.3	7.1/1.3	8.1/1.3	10.1/1.4	12.1/1.5	C	C
Q5_K_S	off	header	3.3/0.5	3.5/0.5	3.8/0.5	4.5/0.6	5.2/0.6	6.0/0.6	7.4/0.6	8.8/0.7	C	C
Q5_K_S	on	header	4.3/1.2	4.5/1.2	5.0/1.2	6.0/1.3	7.0/1.3	8.0/1.3	10.0/1.4	12.0/1.5	C	C
Q6_K	off	header	3.8/0.5	4.0/0.5	4.3/0.5	5.0/0.6	5.7/0.6	6.5/0.6	7.9/0.6	9.3/0.7	C	C
Q6_K	on	header	4.8/1.2	5.0/1.2	5.5/1.2	6.5/1.3	7.5/1.3	8.5/1.3	10.5/1.4	12.5/1.5	C	C
Q8_0	off	file	4.7/0.7	4.9/0.7	5.3/0.7	6.0/0.7	6.7/0.7	7.4/0.8	8.8/0.8	10.3/0.8	C	C
Q8_0	on	file	5.7/1.4	6.0/1.4	6.5/1.4	7.5/1.4	8.5/1.5	9.5/1.5	11.4/1.6	13.4/1.6	C	C
UD-IQ2_M	off	header	2.0/0.5	2.2/0.5	2.5/0.5	3.3/0.5	4.0/0.5	4.7/0.5	6.1/0.5	7.5/0.6	C	C
UD-IQ2_M	on	header	3.0/1.1	3.3/1.1	3.8/1.2	4.8/1.2	5.7/1.2	6.7/1.3	8.7/1.3	10.7/1.4	C	C
UD-IQ2_XXS	off	header	1.8/0.5	2.0/0.5	2.3/0.5	3.0/0.5	3.7/0.5	4.4/0.5	5.9/0.5	7.3/0.6	C	C
UD-IQ2_XXS	on	header	2.8/1.1	3.0/1.1	3.5/1.2	4.5/1.2	5.5/1.2	6.5/1.3	8.5/1.3	10.5/1.4	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.5 4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ3_XXS	off	header	2.2/0.5	2.4/0.5	2.7/0.5	3.5/0.5	4.2/0.5	4.9/0.5	6.3/0.5	7.7/0.6	C	C
UD-IQ3_XXS	on	header	3.2/1.1	3.5/1.1	4.0/1.2	4.9/1.2	5.9/1.2	6.9/1.3	8.9/1.3	10.9/1.4	C	C
UD-Q2_K_XL	off	header	2.2/0.5	2.4/0.5	2.7/0.5	3.4/0.6	4.2/0.6	4.9/0.6	6.3/0.6	7.7/0.7	C	C
UD-Q2_K_XL	on	header	3.2/1.2	3.5/1.2	3.9/1.2	4.9/1.3	5.9/1.3	6.9/1.3	8.9/1.4	10.9/1.5	C	C
UD-Q3_K_XL	off	header	2.7/0.5	2.9/0.5	3.2/0.5	3.9/0.6	4.7/0.6	5.4/0.6	6.8/0.6	8.2/0.7	C	C
UD-Q3_K_XL	on	header	3.7/1.2	3.9/1.2	4.4/1.2	5.4/1.3	6.4/1.3	7.4/1.3	9.4/1.4	11.4/1.5	C	C
UD-Q4_K_XL	off	header	3.2/0.5	3.3/0.5	3.7/0.5	4.4/0.6	5.1/0.6	5.8/0.6	7.3/0.6	8.7/0.7	C	C
UD-Q4_K_XL	on	header	4.2/1.2	4.4/1.2	4.9/1.2	5.9/1.3	6.9/1.3	7.9/1.3	9.9/1.4	11.8/1.5	C	C
UD-Q5_K_XL	off	header	3.5/0.5	3.7/0.5	4.0/0.5	4.8/0.6	5.5/0.6	6.2/0.6	7.6/0.6	9.0/0.7	C	C
UD-Q5_K_XL	on	header	4.5/1.2	4.8/1.2	5.3/1.2	6.2/1.3	7.2/1.3	8.2/1.3	10.2/1.4	12.2/1.5	C	C
UD-Q6_K_XL	off	header	4.4/0.7	4.6/0.7	4.9/0.7	5.6/0.7	6.4/0.7	7.1/0.8	8.5/0.8	9.9/0.8	C	C
UD-Q6_K_XL	on	header	5.4/1.4	5.7/1.4	6.2/1.4	7.1/1.4	8.1/1.5	9.1/1.5	11.1/1.6	13.1/1.6	C	C
UD-Q8_K_XL	off	header	6.2/1.3	6.4/1.3	6.7/1.3	7.5/1.3	8.2/1.3	8.9/1.3	10.3/1.4	11.7/1.4	C	C
UD-Q8_K_XL	on	header	7.2/2.0	7.5/2.0	8.0/2.0	8.9/2.0	9.9/2.1	10.9/2.1	12.9/2.2	14.9/2.2	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Ministral 3 8B Instruct

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	16.6/1.1	17.2/1.1	18.5/1.1	21.1/1.1	23.6/1.1	26.2/1.2	31.3/1.2	36.5/1.2	C	C
BF16	on	header	H	H	H	H	H	H	H	H	C	C
IQ4_NL	off	header	5.3/0.3	5.9/0.3	7.2/0.3	9.8/0.4	12.4/0.4	14.9/0.4	20.1/0.4	25.2/0.5	C	C
IQ4_NL	on	header	H	H	H	H	H	H	H	H	C	C
IQ4_XS	off	header	5.1/0.3	5.7/0.3	7.0/0.3	9.6/0.4	12.1/0.4	14.7/0.4	19.8/0.4	25.0/0.5	C	C
IQ4_XS	on	header	H	H	H	H	H	H	H	H	C	C
Q2_K	off	header	3.8/0.2	4.5/0.2	5.8/0.2	8.3/0.2	10.9/0.2	13.5/0.3	18.6/0.3	23.7/0.3	C	C
Q2_K	on	header	H	H	H	H	H	H	H	H	C	C
Q2_K_L	off	header	3.8/0.3	4.5/0.3	5.8/0.3	8.3/0.4	10.9/0.4	13.5/0.4	18.6/0.4	23.7/0.5	C	C
Q2_K_L	on	header	H	H	H	H	H	H	H	H	C	C
Q3_K_M	off	header	4.7/0.3	5.3/0.3	6.6/0.3	9.2/0.3	11.7/0.3	14.3/0.3	19.4/0.3	24.6/0.4	C	C
Q3_K_M	on	header	H	H	H	H	H	H	H	H	C	C
Q3_K_S	off	header	4.3/0.3	4.9/0.3	6.2/0.3	8.8/0.3	11.4/0.3	13.9/0.3	19.1/0.3	24.2/0.4	C	C
Q3_K_S	on	header	H	H	H	H	H	H	H	H	C	C
Q4_0	off	header	5.3/0.3	5.9/0.3	7.2/0.3	9.8/0.4	12.4/0.4	14.9/0.4	20.1/0.4	25.2/0.5	C	C

Memory matrix / Ministral 3 8B Instruct

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_0	on	header	H	H	H	H	H	H	H	H	C	C
Q4_1	off	header	5.8/0.4	6.4/0.4	7.7/0.4	10.2/0.4	12.8/0.4	15.4/0.4	20.5/0.5	25.6/0.5	C	C
Q4_1	on	header	H	H	H	H	H	H	H	H	C	C
Q4_K_M	off	file	5.6/0.3	6.2/0.3	7.5/0.3	10.1/0.4	12.6/0.4	15.2/0.4	20.3/0.4	25.5/0.5	C	C
Q4_K_M	on	file	H	H	H	H	H	H	H	H	C	C
Q4_K_S	off	header	5.3/0.3	6.0/0.3	7.2/0.3	9.8/0.4	12.4/0.4	14.9/0.4	20.1/0.4	25.2/0.5	C	C
Q4_K_S	on	header	H	H	H	H	H	H	H	H	C	C
Q5_K_M	off	header	6.4/0.4	7.0/0.4	8.3/0.4	10.8/0.4	13.4/0.4	16.0/0.5	21.1/0.5	26.2/0.5	C	C
Q5_K_M	on	header	H	H	H	H	H	H	H	H	C	C
Q5_K_S	off	header	6.2/0.4	6.9/0.4	8.1/0.4	10.7/0.4	13.3/0.4	15.8/0.5	21.0/0.5	26.1/0.5	C	C
Q5_K_S	on	header	H	H	H	H	H	H	H	H	C	C
Q6_K	off	header	7.2/0.5	7.8/0.5	9.1/0.5	11.7/0.5	14.3/0.5	16.8/0.5	22.0/0.6	27.1/0.6	C	C
Q6_K	on	header	H	H	H	H	H	H	H	H	C	C
Q8_0	off	file	9.1/0.6	9.8/0.6	11.1/0.6	13.6/0.6	16.2/0.6	18.8/0.7	23.9/0.7	29.0/0.7	C	C
Q8_0	on	file	H	H	H	H	H	H	H	H	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Ministral 3 8B Instruct

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ1_M	off	header	2.8/0.3	3.5/0.3	4.8/0.3	7.3/0.3	9.9/0.3	12.5/0.3	17.6/0.3	22.7/0.4	C	C
UD-IQ1_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ1_S	off	header	2.7/0.3	3.4/0.3	4.6/0.3	7.2/0.3	9.8/0.3	12.3/0.3	17.5/0.3	22.6/0.4	C	C
UD-IQ1_S	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ2_M	off	header	3.6/0.3	4.2/0.3	5.5/0.3	8.1/0.3	10.7/0.3	13.2/0.3	18.4/0.3	23.5/0.4	C	C
UD-IQ2_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ2_XXS	off	header	3.1/0.3	3.7/0.3	5.0/0.3	7.6/0.3	10.1/0.3	12.7/0.3	17.8/0.3	23.0/0.4	C	C
UD-IQ2_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ3_XXS	off	header	3.9/0.3	4.6/0.3	5.9/0.3	8.4/0.3	11.0/0.3	13.6/0.3	18.7/0.3	23.8/0.4	C	C
UD-IQ3_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q2_K_XL	off	header	3.9/0.3	4.6/0.3	5.9/0.3	8.4/0.4	11.0/0.4	13.6/0.4	18.7/0.4	23.8/0.5	C	C
UD-Q2_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q3_K_XL	off	header	4.8/0.3	5.4/0.3	6.7/0.3	9.3/0.4	11.9/0.4	14.4/0.4	19.6/0.4	24.7/0.5	C	C
UD-Q3_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q4_K_XL	off	header	5.7/0.3	6.3/0.3	7.6/0.3	10.1/0.4	12.7/0.4	15.3/0.4	20.4/0.4	25.5/0.5	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Ministral 3 8B Instruct

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q4_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q5_K_XL	off	header	6.4/0.4	7.0/0.4	8.3/0.4	10.9/0.4	13.4/0.4	16.0/0.5	21.1/0.5	26.3/0.5	C	C
UD-Q5_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q6_K_XL	off	header	7.8/0.6	8.5/0.6	9.7/0.6	12.3/0.6	14.9/0.6	17.4/0.7	22.6/0.7	27.7/0.7	C	C
UD-Q6_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q8_K_XL	off	header	10.7/1.1	11.3/1.1	12.6/1.1	15.2/1.1	17.7/1.1	20.3/1.2	25.4/1.2	30.6/1.2	C	C
UD-Q8_K_XL	on	header	H	H	H	H	H	H	H	H	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 12B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	24.2/2.0	24.3/2.0	24.5/2.0	24.8/2.1	25.2/2.1	25.6/2.1	26.3/2.1	27.0/2.2	C	C
BF16	on	header	24.7/2.3	24.8/2.3	25.0/2.4	25.5/2.4	25.9/2.4	26.4/2.5	27.2/2.5	28.1/2.6	C	C
IQ4_NL	off	header	7.1/0.6	7.2/0.6	7.4/0.6	7.7/0.6	8.1/0.6	8.4/0.6	9.2/0.7	9.9/0.7	C	C
IQ4_NL	on	header	7.6/0.9	7.7/0.9	7.9/0.9	8.4/0.9	8.8/1.0	9.2/1.0	10.1/1.1	11.0/1.1	C	C
IQ4_XS	off	header	6.7/0.6	6.8/0.6	7.0/0.6	7.4/0.6	7.7/0.6	8.1/0.6	8.8/0.7	9.5/0.7	C	C
IQ4_XS	on	header	7.3/0.9	7.4/0.9	7.6/0.9	8.0/0.9	8.5/1.0	8.9/1.0	9.8/1.1	10.6/1.1	C	C
Q3_K_M	off	header	6.1/0.5	6.2/0.5	6.3/0.5	6.7/0.5	7.1/0.5	7.4/0.5	8.1/0.5	8.9/0.6	C	C
Q3_K_M	on	header	6.6/0.8	6.7/0.8	6.9/0.8	7.3/0.8	7.8/0.8	8.2/0.9	9.1/0.9	10.0/1.0	C	C
Q3_K_S	off	header	5.5/0.5	5.6/0.5	5.8/0.5	6.1/0.5	6.5/0.5	6.9/0.5	7.6/0.5	8.3/0.6	C	C
Q3_K_S	on	header	6.0/0.8	6.1/0.8	6.3/0.8	6.8/0.8	7.2/0.8	7.7/0.9	8.5/0.9	9.4/1.0	C	C
Q4_0	off	header	7.6/1.1	7.7/1.1	7.9/1.1	8.2/1.1	8.6/1.1	8.9/1.2	9.7/1.2	10.4/1.2	C	C
Q4_0	on	header	8.1/1.4	8.2/1.4	8.4/1.4	8.9/1.4	9.3/1.5	9.7/1.5	10.6/1.6	11.5/1.6	C	C
Q4_1	off	header	7.8/0.6	7.9/0.6	8.0/0.6	8.4/0.6	8.8/0.6	9.1/0.6	9.8/0.7	10.6/0.7	C	C
Q4_1	on	header	8.3/0.9	8.4/0.9	8.6/0.9	9.0/0.9	9.5/1.0	9.9/1.0	10.8/1.1	11.7/1.1	C	C
Q4_K_M	off	file	7.5/0.6	7.6/0.6	7.8/0.6	8.1/0.6	8.5/0.6	8.8/0.6	9.6/0.7	10.3/0.7	C	C

Memory matrix / Gemma 4 12B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_K_M	on	file	8.0/0.9	8.1/0.9	8.3/0.9	8.8/0.9	9.2/1.0	9.6/1.0	10.5/1.1	11.4/1.1	C	C
Q4_K_S	off	header	7.1/0.6	7.2/0.6	7.4/0.6	7.8/0.6	8.1/0.6	8.5/0.6	9.2/0.7	9.9/0.7	C	C
Q4_K_S	on	header	7.6/0.9	7.8/0.9	8.0/0.9	8.4/0.9	8.8/1.0	9.3/1.0	10.2/1.1	11.0/1.1	C	C
Q5_K_M	off	header	8.8/0.7	8.9/0.7	9.1/0.7	9.4/0.7	9.8/0.8	10.1/0.8	10.9/0.8	11.6/0.8	C	C
Q5_K_M	on	header	9.3/1.0	9.4/1.0	9.6/1.0	10.1/1.1	10.5/1.1	10.9/1.1	11.8/1.2	12.7/1.3	C	C
Q5_K_S	off	header	8.6/0.7	8.7/0.7	8.8/0.7	9.2/0.7	9.6/0.8	9.9/0.8	10.6/0.8	11.4/0.8	C	C
Q5_K_S	on	header	9.1/1.0	9.2/1.0	9.4/1.0	9.9/1.1	10.3/1.1	10.7/1.1	11.6/1.2	12.5/1.3	C	C
Q6_K	off	header	10.2/0.8	10.2/0.9	10.4/0.9	10.8/0.9	11.1/0.9	11.5/0.9	12.2/0.9	12.9/1.0	C	C
Q6_K	on	header	10.7/1.1	10.8/1.2	11.0/1.2	11.4/1.2	11.9/1.2	12.3/1.3	13.2/1.3	14.1/1.4	C	C
Q8_0	off	file	13.0/1.1	13.1/1.1	13.3/1.1	13.7/1.1	14.0/1.1	14.4/1.2	15.1/1.2	15.8/1.2	C	C
Q8_0	on	file	13.6/1.4	13.7/1.4	13.9/1.4	14.3/1.4	14.8/1.5	15.2/1.5	16.1/1.6	16.9/1.6	C	C
UD-IQ2_M	off	header	4.6/0.5	4.7/0.5	4.9/0.5	5.2/0.5	5.6/0.5	5.9/0.5	6.7/0.5	7.4/0.6	C	C
UD-IQ2_M	on	header	5.1/0.8	5.2/0.8	5.4/0.8	5.9/0.8	6.3/0.8	6.7/0.9	7.6/0.9	8.5/1.0	C	C
UD-IQ3_XXS	off	header	5.0/0.5	5.1/0.5	5.3/0.5	5.6/0.5	6.0/0.5	6.4/0.5	7.1/0.5	7.8/0.6	C	C
UD-IQ3_XXS	on	header	5.5/0.8	5.6/0.8	5.9/0.8	6.3/0.8	6.7/0.8	7.2/0.9	8.0/0.9	8.9/1.0	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 12B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q2_K_XL	off	header	5.0/0.6	5.1/0.6	5.3/0.6	5.7/0.6	6.0/0.6	6.4/0.6	7.1/0.7	7.8/0.7	C	C
UD-Q2_K_XL	on	header	5.5/0.9	5.7/0.9	5.9/0.9	6.3/0.9	6.7/1.0	7.2/1.0	8.1/1.1	8.9/1.1	C	C
UD-Q3_K_XL	off	header	6.4/0.6	6.5/0.6	6.7/0.6	7.0/0.6	7.4/0.6	7.7/0.6	8.5/0.7	9.2/0.7	C	C
UD-Q3_K_XL	on	header	6.9/0.9	7.0/0.9	7.2/0.9	7.7/0.9	8.1/1.0	8.5/1.0	9.4/1.1	10.3/1.1	C	C
UD-Q4_K_XL	off	header	7.7/0.7	7.8/0.7	8.0/0.7	8.4/0.7	8.7/0.8	9.1/0.8	9.8/0.8	10.5/0.8	C	C
UD-Q4_K_XL	on	header	8.3/1.0	8.4/1.0	8.6/1.0	9.0/1.1	9.4/1.1	9.9/1.1	10.8/1.2	11.6/1.3	C	C
UD-Q5_K_XL	off	header	9.0/0.8	9.1/0.9	9.2/0.9	9.6/0.9	10.0/0.9	10.3/0.9	11.0/0.9	11.8/1.0	C	C
UD-Q5_K_XL	on	header	9.5/1.1	9.6/1.2	9.8/1.2	10.3/1.2	10.7/1.2	11.1/1.3	12.0/1.3	12.9/1.4	C	C
UD-Q6_K_XL	off	header	11.1/1.1	11.1/1.1	11.3/1.1	11.7/1.1	12.0/1.1	12.4/1.2	13.1/1.2	13.8/1.2	C	C
UD-Q6_K_XL	on	header	11.6/1.4	11.7/1.4	11.9/1.4	12.3/1.4	12.8/1.5	13.2/1.5	14.1/1.6	14.9/1.6	C	C
UD-Q8_K_XL	off	header	14.0/1.1	14.1/1.1	14.3/1.1	14.6/1.1	15.0/1.1	15.4/1.2	16.1/1.2	16.8/1.2	C	C
UD-Q8_K_XL	on	header	14.5/1.4	14.6/1.4	14.8/1.4	15.3/1.4	15.7/1.5	16.2/1.5	17.0/1.6	17.9/1.6	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.8 27B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	51.8/2.6	52.1/2.6	52.7/2.6	54.0/2.6	55.3/2.6	56.6/2.6	59.2/2.7	61.7/2.7	C	C
BF16	on	header	54.0/3.9	54.4/3.9	55.2/4.0	56.7/4.0	58.3/4.0	59.9/4.1	63.0/4.1	66.1/4.2	C	C
Q4_0	off	header	15.6/0.7	15.9/0.7	16.6/0.8	17.8/0.8	19.1/0.8	20.4/0.8	23.0/0.8	25.5/0.9	C	C
Q4_0	on	header	16.3/0.8	16.7/0.8	17.5/0.8	19.0/0.8	20.6/0.8	22.1/0.9	25.3/0.9	28.4/1.0	C	C
Q4_1	off	header	17.0/0.8	17.3/0.8	17.9/0.8	19.2/0.8	20.5/0.9	21.8/0.9	24.4/0.9	26.9/0.9	C	C
Q4_1	on	header	17.7/0.8	18.1/0.8	18.9/0.9	20.4/0.9	22.0/0.9	23.5/1.0	26.7/1.0	29.8/1.1	C	C
Q4_K_M	off	header	19.2/0.7	19.5/0.7	20.2/0.8	21.5/0.8	22.7/0.8	24.0/0.8	26.6/0.8	29.2/0.9	C	C
Q4_K_M	off	file	16.3/0.7	16.7/0.7	17.3/0.8	18.6/0.8	19.9/0.8	21.2/0.8	23.7/0.8	26.3/0.9	C	C
Q4_K_M	on	header	21.4/2.1	21.8/2.1	22.6/2.1	24.2/2.2	25.7/2.2	27.3/2.2	30.4/2.3	33.5/2.4	C	C
Q4_K_M	on	file	17.0/0.8	17.4/0.8	18.2/0.8	19.8/0.8	21.3/0.8	22.9/0.9	26.0/0.9	29.1/1.0	C	C
Q8_0	off	file	27.8/1.4	28.1/1.4	28.7/1.4	30.0/1.4	31.3/1.4	32.6/1.4	35.1/1.5	37.7/1.5	C	C
Q8_0	on	file	30.0/2.7	30.4/2.7	31.2/2.8	32.7/2.8	34.3/2.8	35.8/2.9	39.0/2.9	42.1/3.0	C	C
UD-IQ1_M	off	header	6.8/0.4	7.1/0.4	7.8/0.5	9.1/0.5	10.4/0.5	11.6/0.5	14.2/0.5	16.8/0.6	C	C
UD-IQ1_M	on	header	9.1/1.8	9.5/1.8	10.2/1.8	11.8/1.9	13.4/1.9	14.9/1.9	18.0/2.0	21.2/2.1	C	C
UD-IQ1_S	off	header	6.3/0.4	6.6/0.4	7.3/0.5	8.5/0.5	9.8/0.5	11.1/0.5	13.7/0.5	16.2/0.6	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.8 27B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ1_S	on	header	8.5/1.8	8.9/1.8	9.7/1.8	11.3/1.9	12.8/1.9	14.4/1.9	17.5/2.0	20.6/2.1	C	C
UD-IQ2_S	off	header	8.5/0.4	8.8/0.4	9.5/0.5	10.8/0.5	12.1/0.5	13.3/0.5	15.9/0.5	18.5/0.6	C	C
UD-IQ2_S	on	header	10.8/1.8	11.1/1.8	11.9/1.8	13.5/1.9	15.0/1.9	16.6/1.9	19.7/2.0	22.8/2.1	C	C
UD-IQ2_XS	off	header	7.8/0.4	8.1/0.4	8.8/0.4	10.1/0.4	11.3/0.4	12.6/0.4	15.2/0.4	17.8/0.5	C	C
UD-IQ2_XS	on	header	10.0/1.7	10.4/1.7	11.2/1.7	12.8/1.8	14.3/1.8	15.9/1.8	19.0/1.9	22.1/2.0	C	C
UD-IQ2_XXS	off	header	7.4/0.4	7.7/0.4	8.4/0.4	9.6/0.4	10.9/0.4	12.2/0.4	14.8/0.4	17.3/0.5	C	C
UD-IQ2_XXS	on	header	9.6/1.7	10.0/1.7	10.8/1.7	12.4/1.8	13.9/1.8	15.5/1.8	18.6/1.9	21.7/2.0	C	C
UD-IQ3_S	off	header	11.7/0.6	12.0/0.6	12.6/0.6	13.9/0.6	15.2/0.6	16.5/0.6	19.0/0.7	21.6/0.7	C	C
UD-IQ3_S	on	header	12.4/0.6	12.8/0.6	13.6/0.6	15.2/0.6	16.7/0.7	18.3/0.7	21.4/0.8	24.5/0.8	C	C
UD-IQ3_XXS	off	header	10.7/0.4	11.0/0.4	11.6/0.5	12.9/0.5	14.2/0.5	15.5/0.5	18.1/0.5	20.6/0.6	C	C
UD-IQ3_XXS	on	header	11.5/0.5	11.9/0.5	12.6/0.5	14.2/0.5	15.8/0.5	17.3/0.6	20.4/0.6	23.6/0.7	C	C
UD-IQ4_XS	off	header	14.0/0.6	14.3/0.6	14.9/0.6	16.2/0.6	17.5/0.6	18.8/0.6	21.4/0.7	23.9/0.7	C	C
UD-IQ4_XS	on	header	14.8/0.6	15.1/0.6	15.9/0.6	17.5/0.6	19.0/0.7	20.6/0.7	23.7/0.8	26.8/0.8	C	C
UD-Q2_K_XL	off	header	9.7/0.4	10.0/0.4	10.7/0.5	11.9/0.5	13.2/0.5	14.5/0.5	17.1/0.5	19.6/0.6	C	C
UD-Q2_K_XL	on	header	10.5/0.5	10.9/0.5	11.6/0.5	13.2/0.5	14.8/0.5	16.3/0.6	19.4/0.6	22.6/0.7	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.8 27B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q3_K_XL	off	header	12.8/0.6	13.1/0.6	13.7/0.6	15.0/0.6	16.3/0.6	17.6/0.6	20.1/0.7	22.7/0.7	C	C
UD-Q3_K_XL	on	header	13.5/0.6	13.9/0.6	14.7/0.6	16.3/0.6	17.8/0.7	19.4/0.7	22.5/0.8	25.6/0.8	C	C
UD-Q4_K_M	off	header	16.0/0.7	16.3/0.7	17.0/0.8	18.3/0.8	19.5/0.8	20.8/0.8	23.4/0.8	26.0/0.9	C	C
UD-Q4_K_M	on	header	16.8/0.8	17.2/0.8	18.0/0.8	19.5/0.8	21.1/0.8	22.6/0.9	25.8/0.9	28.9/1.0	C	C
UD-Q4_K_S	off	header	15.0/0.6	15.3/0.6	15.9/0.6	17.2/0.6	18.5/0.6	19.8/0.6	22.4/0.7	24.9/0.7	C	C
UD-Q4_K_S	on	header	15.8/0.6	16.1/0.6	16.9/0.6	18.5/0.6	20.0/0.7	21.6/0.7	24.7/0.8	27.9/0.8	C	C
UD-Q4_K_XL	off	header	17.1/0.7	17.4/0.7	18.1/0.8	19.4/0.8	20.6/0.8	21.9/0.8	24.5/0.8	27.1/0.9	C	C
UD-Q4_K_XL	on	header	17.9/0.8	18.3/0.8	19.1/0.8	20.6/0.8	22.2/0.8	23.7/0.9	26.9/0.9	30.0/1.0	C	C
UD-Q5_K_M	off	header	19.3/0.7	19.6/0.7	20.3/0.8	21.6/0.8	22.9/0.8	24.1/0.8	26.7/0.8	29.3/0.9	C	C
UD-Q5_K_M	on	header	20.1/0.8	20.5/0.8	21.3/0.8	22.8/0.8	24.4/0.8	26.0/0.9	29.1/0.9	32.2/1.0	C	C
UD-Q5_K_S	off	header	18.2/0.7	18.5/0.7	19.2/0.8	20.5/0.8	21.7/0.8	23.0/0.8	25.6/0.8	28.2/0.9	C	C
UD-Q5_K_S	on	header	19.0/0.8	19.4/0.8	20.2/0.8	21.7/0.8	23.3/0.8	24.8/0.9	28.0/0.9	31.1/1.0	C	C
UD-Q5_K_XL	off	header	20.1/1.1	20.4/1.1	21.1/1.1	22.3/1.1	23.6/1.1	24.9/1.1	27.5/1.2	30.0/1.2	C	C
UD-Q5_K_XL	on	header	20.9/1.1	21.3/1.1	22.1/1.1	23.6/1.1	25.2/1.2	26.7/1.2	29.9/1.3	33.0/1.3	C	C
UD-Q6_K	off	header	21.1/1.1	21.4/1.1	22.1/1.1	23.3/1.1	24.6/1.1	25.9/1.1	28.5/1.2	31.1/1.2	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.8 27B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q6_K	on	header	21.9/1.1	22.3/1.1	23.1/1.1	24.6/1.1	26.2/1.2	27.7/1.2	30.9/1.3	34.0/1.3	C	C
UD-Q6_K_L	off	header	23.3/1.1	23.6/1.1	24.3/1.1	25.6/1.1	26.8/1.1	28.1/1.1	30.7/1.2	33.3/1.2	C	C
UD-Q6_K_L	on	header	24.1/1.1	24.5/1.1	25.3/1.1	26.8/1.1	28.4/1.2	29.9/1.2	33.1/1.3	36.2/1.3	C	C
UD-Q6_K_M	off	header	22.2/1.1	22.5/1.1	23.2/1.1	24.5/1.1	25.7/1.1	27.0/1.1	29.6/1.2	32.2/1.2	C	C
UD-Q6_K_M	on	header	23.0/1.1	23.4/1.1	24.2/1.1	25.7/1.1	27.3/1.2	28.8/1.2	32.0/1.3	35.1/1.3	C	C
UD-Q6_K_XL	off	header	24.4/1.1	24.7/1.1	25.4/1.1	26.7/1.1	28.0/1.1	29.2/1.1	31.8/1.2	34.4/1.2	C	C
UD-Q6_K_XL	on	header	25.2/1.1	25.6/1.1	26.4/1.1	27.9/1.1	29.5/1.2	31.1/1.2	34.2/1.3	37.3/1.3	C	C
UD-Q8_K_L	off	header	26.9/1.4	27.2/1.4	27.8/1.4	29.1/1.4	30.4/1.4	31.7/1.4	34.2/1.5	36.8/1.5	C	C
UD-Q8_K_L	on	header	27.6/1.4	28.0/1.4	28.8/1.4	30.4/1.4	31.9/1.5	33.5/1.5	36.6/1.6	39.7/1.7	C	C
UD-Q8_K_XL	off	header	30.1/1.4	30.4/1.4	31.0/1.4	32.3/1.4	33.6/1.4	34.9/1.4	37.4/1.5	40.0/1.5	C	C
UD-Q8_K_XL	on	header	31.1/1.4	31.4/1.4	32.2/1.4	33.8/1.4	35.3/1.5	36.9/1.5	40.0/1.6	43.1/1.7	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 31B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	62.5/2.8	62.9/2.8	63.8/2.9	65.5/2.9	67.2/2.9	68.9/2.9	72.3/2.9	75.7/3.0	C	C
BF16	on	header	63.2/3.1	63.7/3.2	64.6/3.2	66.6/3.2	68.6/3.2	70.6/3.3	74.5/3.3	78.5/3.4	C	C
IQ4_NL	off	header	18.4/0.8	18.8/0.8	19.7/0.8	21.4/0.8	23.1/0.9	24.8/0.9	28.2/0.9	31.6/0.9	C	C
IQ4_NL	on	header	19.0/1.1	19.5/1.1	20.5/1.1	22.5/1.2	24.5/1.2	26.5/1.2	30.4/1.3	34.4/1.4	C	C
IQ4_XS	off	header	17.5/0.8	17.9/0.8	18.7/0.8	20.4/0.8	22.1/0.9	23.8/0.9	27.3/0.9	30.7/0.9	C	C
IQ4_XS	on	header	18.1/1.1	18.6/1.1	19.6/1.1	21.6/1.2	23.6/1.2	25.5/1.2	29.5/1.3	33.5/1.4	C	C
Q3_K_M	off	header	15.8/0.6	16.2/0.6	17.1/0.6	18.8/0.7	20.5/0.7	22.2/0.7	25.6/0.7	29.0/0.8	C	C
Q3_K_M	on	header	16.5/0.9	17.0/0.9	18.0/1.0	19.9/1.0	21.9/1.0	23.9/1.1	27.9/1.1	31.8/1.2	C	C
Q3_K_S	off	header	14.3/0.6	14.7/0.6	15.6/0.6	17.3/0.7	19.0/0.7	20.7/0.7	24.1/0.7	27.5/0.8	C	C
Q3_K_S	on	header	15.0/0.9	15.5/0.9	16.4/1.0	18.4/1.0	20.4/1.0	22.4/1.1	26.3/1.1	30.3/1.2	C	C
Q4_0	off	header	19.1/1.5	19.5/1.5	20.4/1.5	22.1/1.6	23.8/1.6	25.5/1.6	28.9/1.6	32.3/1.7	C	C
Q4_0	on	header	19.7/1.8	20.2/1.8	21.2/1.9	23.2/1.9	25.2/1.9	27.2/2.0	31.1/2.0	35.1/2.1	C	C
Q4_1	off	header	20.2/0.8	20.6/0.8	21.5/0.8	23.2/0.8	24.9/0.9	26.6/0.9	30.0/0.9	33.4/0.9	C	C
Q4_1	on	header	20.9/1.1	21.4/1.1	22.4/1.1	24.3/1.2	26.3/1.2	28.3/1.2	32.2/1.3	36.2/1.4	C	C
Q4_K_M	off	file	19.4/0.8	19.8/0.8	20.7/0.8	22.4/0.8	24.1/0.9	25.8/0.9	29.2/0.9	32.6/0.9	C	C

Memory matrix / Gemma 4 31B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_K_M	on	file	20.1/1.1	20.6/1.1	21.6/1.1	23.5/1.2	25.5/1.2	27.5/1.2	31.5/1.3	35.4/1.4	C	C
Q4_K_S	off	header	18.5/0.8	18.9/0.8	19.8/0.8	21.5/0.8	23.2/0.9	24.9/0.9	28.3/0.9	31.7/0.9	C	C
Q4_K_S	on	header	19.1/1.1	19.6/1.1	20.6/1.1	22.6/1.2	24.6/1.2	26.6/1.2	30.5/1.3	34.5/1.4	C	C
Q5_K_M	off	header	22.7/1.0	23.2/1.0	24.0/1.0	25.7/1.0	27.4/1.0	29.1/1.1	32.5/1.1	35.9/1.1	C	C
Q5_K_M	on	header	23.4/1.3	23.9/1.3	24.9/1.3	26.9/1.4	28.9/1.4	30.8/1.4	34.8/1.5	38.7/1.6	C	C
Q5_K_S	off	header	22.2/1.0	22.6/1.0	23.5/1.0	25.2/1.0	26.9/1.0	28.6/1.1	32.0/1.1	35.4/1.1	C	C
Q5_K_S	on	header	22.9/1.3	23.4/1.3	24.4/1.3	26.3/1.4	28.3/1.4	30.3/1.4	34.3/1.5	38.2/1.6	C	C
Q6_K	off	header	26.3/1.2	26.7/1.2	27.6/1.2	29.3/1.2	31.0/1.2	32.7/1.2	36.1/1.3	39.5/1.3	C	C
Q6_K	on	header	26.9/1.5	27.4/1.5	28.4/1.5	30.4/1.5	32.4/1.6	34.4/1.6	38.3/1.7	42.3/1.7	C	C
Q8_0	off	file	33.7/1.5	34.1/1.5	35.0/1.5	36.7/1.6	38.4/1.6	40.1/1.6	43.5/1.6	46.9/1.7	C	C
Q8_0	on	file	34.4/1.8	34.9/1.8	35.9/1.9	37.8/1.9	39.8/1.9	41.8/2.0	45.8/2.0	49.7/2.1	C	C
UD-IQ2_M	off	header	11.8/0.6	12.3/0.6	13.1/0.6	14.8/0.7	16.5/0.7	18.2/0.7	21.6/0.7	25.0/0.8	C	C
UD-IQ2_M	on	header	12.5/0.9	13.0/0.9	14.0/1.0	16.0/1.0	17.9/1.0	19.9/1.1	23.9/1.1	27.8/1.2	C	C
UD-IQ2_XXS	off	header	9.6/0.6	10.0/0.6	10.9/0.6	12.6/0.7	14.3/0.7	16.0/0.7	19.4/0.7	22.8/0.8	C	C
UD-IQ2_XXS	on	header	10.3/0.9	10.8/0.9	11.8/1.0	13.7/1.0	15.7/1.0	17.7/1.1	21.7/1.1	25.6/1.2	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 31B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ3_XXS	off	header	12.9/0.6	13.4/0.6	14.2/0.6	15.9/0.7	17.6/0.7	19.3/0.7	22.7/0.7	26.1/0.8	C	C
UD-IQ3_XXS	on	header	13.6/0.9	14.1/0.9	15.1/1.0	17.1/1.0	19.0/1.0	21.0/1.1	25.0/1.1	28.9/1.2	C	C
UD-Q2_K_XL	off	header	12.9/0.8	13.3/0.8	14.1/0.8	15.8/0.8	17.5/0.9	19.2/0.9	22.7/0.9	26.1/0.9	C	C
UD-Q2_K_XL	on	header	13.5/1.1	14.0/1.1	15.0/1.1	17.0/1.2	19.0/1.2	20.9/1.2	24.9/1.3	28.9/1.4	C	C
UD-Q3_K_XL	off	header	16.5/0.8	16.9/0.8	17.7/0.8	19.4/0.8	21.2/0.9	22.9/0.9	26.3/0.9	29.7/0.9	C	C
UD-Q3_K_XL	on	header	17.1/1.1	17.6/1.1	18.6/1.1	20.6/1.2	22.6/1.2	24.6/1.2	28.5/1.3	32.5/1.4	C	C
UD-Q4_K_XL	off	header	19.9/1.0	20.3/1.0	21.2/1.0	22.9/1.0	24.6/1.0	26.3/1.1	29.7/1.1	33.1/1.1	C	C
UD-Q4_K_XL	on	header	20.6/1.3	21.1/1.3	22.1/1.3	24.0/1.4	26.0/1.4	28.0/1.4	32.0/1.5	35.9/1.6	C	C
UD-Q5_K_XL	off	header	23.0/1.2	23.4/1.2	24.3/1.2	26.0/1.2	27.7/1.2	29.4/1.2	32.8/1.3	36.2/1.3	C	C
UD-Q5_K_XL	on	header	23.6/1.5	24.1/1.5	25.1/1.5	27.1/1.5	29.1/1.6	31.1/1.6	35.0/1.7	39.0/1.7	C	C
UD-Q6_K_XL	off	header	28.6/1.5	29.0/1.5	29.9/1.5	31.6/1.6	33.3/1.6	35.0/1.6	38.4/1.6	41.8/1.7	C	C
UD-Q6_K_XL	on	header	29.3/1.8	29.8/1.8	30.8/1.9	32.7/1.9	34.7/1.9	36.7/2.0	40.7/2.0	44.6/2.1	C	C
UD-Q8_K_XL	off	header	36.1/1.5	36.5/1.5	37.4/1.5	39.1/1.6	40.8/1.6	42.5/1.6	45.9/1.6	49.3/1.7	C	C
UD-Q8_K_XL	on	header	36.8/1.8	37.3/1.8	38.3/1.9	40.2/1.9	42.2/1.9	44.2/2.0	48.2/2.0	52.1/2.1	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Inkling 2 Mini

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	51.8/0.9	52.2/1.0	52.9/1.0	54.3/1.0	55.7/1.1	57.1/1.1	60.0/1.2	62.8/1.3	74.1/1.7	96.8/2.5
BF16	on	header	H	H	H	H	H	H	H	H	H	H
IQ4_NL	off	file	17.6/0.5	18.0/0.5	18.7/0.5	20.1/0.6	21.5/0.6	23.0/0.7	25.8/0.8	28.6/0.9	40.0/1.3	62.7/2.1
IQ4_NL	on	file	H	H	H	H	H	H	H	H	H	H
Q8_0	off	file	27.8/0.5	28.2/0.5	28.9/0.5	30.3/0.6	31.7/0.6	33.2/0.7	36.0/0.8	38.8/0.9	50.2/1.3	72.8/2.1
Q8_0	on	file	H	H	H	H	H	H	H	H	H	H
Q8_0_OLD	off	header	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Q8_0_OLD	on	header	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
UD-IQ1_M	off	header	7.7/0.9	8.1/0.9	8.8/0.9	10.2/0.9	11.6/1.0	13.0/1.0	15.9/1.1	18.7/1.2	30.0/1.6	52.7/2.4
UD-IQ1_M	on	header	H	H	H	H	H	H	H	H	H	H
UD-IQ1_S	off	header	7.8/0.5	8.2/0.5	8.9/0.5	10.3/0.5	11.7/0.6	13.2/0.6	16.0/0.7	18.8/0.8	30.2/1.2	52.8/2.1
UD-IQ1_S	on	header	H	H	H	H	H	H	H	H	H	H
UD-IQ2_XXS	off	header	7.4/1.6	7.8/1.6	8.5/1.6	9.9/1.7	11.3/1.7	12.7/1.8	15.6/1.9	18.4/2.0	29.7/2.4	52.4/3.2
UD-IQ2_XXS	on	header	H	H	H	H	H	H	H	H	H	H
UD-IQ3_XXS	off	header	8.1/4.1	8.4/4.1	9.2/4.2	10.6/4.2	12.0/4.3	13.4/4.3	16.2/4.4	19.1/4.5	30.4/4.9	53.1/5.7

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Inkling 2 Mini

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ3_XXS	on	header	H	H	H	H	H	H	H	H	H	H
UD-IQ4_XS	off	header	4.3/12.0	4.6/12.0	5.3/12.1	6.8/12.1	8.2/12.2	9.6/12.2	12.4/12.3	15.3/12.4	26.6/12.8	49.3/13.6
UD-IQ4_XS	on	header	H	H	H	H	H	H	H	H	H	H
UD-Q2_K_XL	off	header	7.3/2.4	7.7/2.4	8.4/2.4	9.8/2.5	11.2/2.5	12.6/2.6	15.4/2.7	18.3/2.8	29.6/3.2	52.3/4.0
UD-Q2_K_XL	on	header	H	H	H	H	H	H	H	H	H	H
UD-Q3_K_XL	off	header	5.4/8.8	5.8/8.8	6.5/8.8	7.9/8.9	9.3/8.9	10.7/9.0	13.6/9.1	16.4/9.2	27.7/9.6	50.4/10.4
UD-Q3_K_XL	on	header	H	H	H	H	H	H	H	H	H	H
UD-Q4_K_M	off	header	5.2/12.0	5.5/12.0	6.2/12.1	7.6/12.1	9.1/12.2	10.5/12.2	13.3/12.3	16.1/12.4	27.5/12.8	50.2/13.6
UD-Q4_K_M	on	header	H	H	H	H	H	H	H	H	H	H
UD-Q4_K_XL	off	header	6.2/12.0	6.5/12.0	7.2/12.1	8.7/12.1	10.1/12.2	11.5/12.2	14.3/12.3	17.2/12.4	28.5/12.8	51.2/13.6
UD-Q4_K_XL	on	header	H	H	H	H	H	H	H	H	H	H
UD-Q8_K_XL	off	header	11.0/12.4	11.3/12.5	12.0/12.5	13.5/12.5	14.9/12.6	16.3/12.6	19.1/12.7	22.0/12.8	33.3/13.2	56.0/14.0
UD-Q8_K_XL	on	header	H	H	H	H	H	H	H	H	H	H

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Muse Glimmer 30B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	53.2/2.7	53.2/2.7	53.4/2.7	53.7/2.8	53.9/2.8	54.2/2.8	C	C	C	C
BF16	on	header	H	H	H	H	H	H	C	C	C	C
IQ2_M	off	header	10.2/0.6	10.3/0.6	10.4/0.6	10.7/0.6	11.0/0.7	11.2/0.7	C	C	C	C
IQ2_M	on	header	H	H	H	H	H	H	C	C	C	C
IQ2_S	off	header	9.6/0.6	9.7/0.6	9.8/0.6	10.1/0.6	10.4/0.7	10.6/0.7	C	C	C	C
IQ2_S	on	header	H	H	H	H	H	H	C	C	C	C
IQ2_XS	off	header	9.3/0.5	9.4/0.5	9.5/0.5	9.8/0.5	10.0/0.5	10.3/0.5	C	C	C	C
IQ2_XS	on	header	H	H	H	H	H	H	C	C	C	C
IQ2_XXS	off	header	8.6/0.5	8.7/0.5	8.8/0.5	9.1/0.5	9.4/0.5	9.6/0.5	C	C	C	C
IQ2_XXS	on	header	H	H	H	H	H	H	C	C	C	C
IQ3_M	off	header	12.7/0.6	12.7/0.6	12.9/0.6	13.2/0.6	13.4/0.7	13.7/0.7	C	C	C	C
IQ3_M	on	header	H	H	H	H	H	H	C	C	C	C
IQ3_XS	off	header	11.9/0.6	11.9/0.6	12.1/0.6	12.4/0.6	12.6/0.7	12.9/0.7	C	C	C	C
IQ3_XS	on	header	H	H	H	H	H	H	C	C	C	C
IQ3_XXS	off	header	11.1/0.6	11.2/0.6	11.3/0.6	11.6/0.6	11.9/0.7	12.1/0.7	C	C	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Muse Glimmer 30B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
IQ3_XXS	on	header	H	H	H	H	H	H	C	C	C	C
IQ4_NL	off	header	15.6/0.8	15.7/0.8	15.8/0.8	16.1/0.8	16.4/0.8	16.7/0.9	C	C	C	C
IQ4_NL	on	header	H	H	H	H	H	H	C	C	C	C
IQ4_XS	off	header	14.9/0.7	14.9/0.8	15.1/0.8	15.3/0.8	15.6/0.8	15.9/0.8	C	C	C	C
IQ4_XS	on	header	H	H	H	H	H	H	C	C	C	C
Q2_K	off	header	10.7/0.5	10.8/0.5	10.9/0.5	11.2/0.5	11.5/0.5	11.8/0.5	C	C	C	C
Q2_K	on	header	H	H	H	H	H	H	C	C	C	C
Q2_K_L	off	header	11.1/1.5	11.1/1.5	11.3/1.5	11.5/1.5	11.8/1.5	12.1/1.5	C	C	C	C
Q2_K_L	on	header	H	H	H	H	H	H	C	C	C	C
Q3_K_L	off	header	14.3/0.6	14.4/0.6	14.6/0.6	14.8/0.6	15.1/0.7	15.4/0.7	C	C	C	C
Q3_K_L	on	header	H	H	H	H	H	H	C	C	C	C
Q3_K_M	off	header	13.5/0.6	13.6/0.6	13.7/0.6	14.0/0.6	14.3/0.7	14.6/0.7	C	C	C	C
Q3_K_M	on	header	H	H	H	H	H	H	C	C	C	C
Q3_K_S	off	header	12.4/0.6	12.4/0.6	12.6/0.6	12.8/0.6	13.1/0.7	13.4/0.7	C	C	C	C
Q3_K_S	on	header	H	H	H	H	H	H	C	C	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Muse Glimmer 30B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q3_K_XL	off	header	14.7/1.5	14.7/1.5	14.9/1.5	15.2/1.5	15.4/1.5	15.7/1.5	C	C	C	C
Q3_K_XL	on	header	H	H	H	H	H	H	C	C	C	C
Q4_0	off	header	15.7/0.8	15.7/0.8	15.9/0.8	16.1/0.8	16.4/0.8	16.7/0.9	C	C	C	C
Q4_0	on	header	H	H	H	H	H	H	C	C	C	C
Q4_1	off	header	17.1/0.9	17.2/0.9	17.3/0.9	17.6/0.9	17.9/0.9	18.2/0.9	C	C	C	C
Q4_1	on	header	H	H	H	H	H	H	C	C	C	C
Q4_K_L	off	header	17.0/1.5	17.1/1.5	17.2/1.5	17.5/1.5	17.8/1.5	18.0/1.5	C	C	C	C
Q4_K_L	on	header	H	H	H	H	H	H	C	C	C	C
Q4_K_M	off	file	16.7/0.8	16.8/0.8	16.9/0.8	17.2/0.8	17.4/0.8	17.7/0.9	C	C	C	C
Q4_K_M	on	file	H	H	H	H	H	H	C	C	C	C
Q4_K_S	off	header	15.7/0.8	15.8/0.8	15.9/0.8	16.2/0.8	16.5/0.8	16.7/0.9	C	C	C	C
Q4_K_S	on	header	H	H	H	H	H	H	C	C	C	C
Q5_K_L	off	header	19.7/1.5	19.7/1.5	19.9/1.5	20.1/1.5	20.4/1.5	20.7/1.5	C	C	C	C
Q5_K_L	on	header	H	H	H	H	H	H	C	C	C	C
Q5_K_M	off	header	19.3/1.0	19.4/1.0	19.5/1.0	19.8/1.0	20.1/1.0	20.4/1.0	C	C	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Muse Glimmer 30B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q5_K_M	on	header	H	H	H	H	H	H	C	C	C	C
Q5_K_S	off	header	18.7/1.0	18.7/1.0	18.9/1.0	19.1/1.0	19.4/1.0	19.7/1.0	C	C	C	C
Q5_K_S	on	header	H	H	H	H	H	H	C	C	C	C
Q6_K	off	header	22.5/1.1	22.5/1.1	22.7/1.1	22.9/1.2	23.2/1.2	23.5/1.2	C	C	C	C
Q6_K	on	header	H	H	H	H	H	H	C	C	C	C
Q6_K_L	off	header	22.8/1.5	22.8/1.5	23.0/1.5	23.3/1.5	23.5/1.5	23.8/1.5	C	C	C	C
Q6_K_L	on	header	H	H	H	H	H	H	C	C	C	C
Q8_0	off	file	28.3/1.5	28.4/1.5	28.5/1.5	28.8/1.5	29.1/1.5	29.4/1.5	C	C	C	C
Q8_0	on	file	H	H	H	H	H	H	C	C	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.6 35B-A3B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	68.6/1.0	68.7/1.0	68.9/1.0	69.3/1.1	69.7/1.1	70.2/1.1	71.0/1.1	71.9/1.2	C	C
BF16	on	header	H	H	H	H	H	H	H	H	C	C
MXFP4_MOE	off	header	21.4/0.6	21.5/0.6	21.7/0.6	22.1/0.6	22.6/0.6	23.0/0.6	23.8/0.6	24.7/0.7	C	C
MXFP4_MOE	on	header	H	H	H	H	H	H	H	H	C	C
Q4_K_M	off	header	20.5/0.3	20.6/0.3	20.8/0.3	21.2/0.3	21.6/0.3	22.1/0.4	22.9/0.4	23.8/0.4	C	C
Q4_K_M	on	header	H	H	H	H	H	H	H	H	C	C
Q8_0	off	header	36.6/0.6	36.7/0.6	36.9/0.6	37.3/0.6	37.7/0.6	38.2/0.6	39.0/0.6	39.9/0.7	C	C
Q8_0	off	file	36.6/0.6	36.7/0.6	36.9/0.6	37.3/0.6	37.7/0.6	38.2/0.6	39.0/0.6	39.9/0.7	C	C
Q8_0	on	header	H	H	H	H	H	H	H	H	C	C
Q8_0	on	file	37.7/0.6	37.8/0.6	38.1/0.6	38.7/0.6	39.2/0.7	39.8/0.7	41.0/0.8	42.1/0.8	C	C
UD-IQ1_M	off	header	10.0/0.3	10.1/0.3	10.3/0.3	10.7/0.3	11.1/0.3	11.6/0.4	12.4/0.4	13.3/0.4	C	C
UD-IQ1_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ2_M	off	header	11.4/0.4	11.5/0.4	11.7/0.4	12.1/0.4	12.6/0.4	13.0/0.4	13.9/0.5	14.7/0.5	C	C
UD-IQ2_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ2_XXS	off	header	10.7/0.3	10.8/0.3	11.0/0.3	11.4/0.3	11.9/0.3	12.3/0.4	13.2/0.4	14.0/0.4	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.6 35B-A3B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ2_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ3_S	off	header	13.5/0.4	13.6/0.4	13.8/0.4	14.2/0.5	14.6/0.5	15.1/0.5	15.9/0.5	16.8/0.6	C	C
UD-IQ3_S	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ3_XXS	off	header	13.0/0.4	13.1/0.4	13.3/0.4	13.7/0.5	14.2/0.5	14.6/0.5	15.5/0.5	16.3/0.6	C	C
UD-IQ3_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ4_NL	off	header	17.7/0.6	17.8/0.6	18.0/0.6	18.5/0.6	18.9/0.6	19.3/0.6	20.2/0.6	21.0/0.7	C	C
UD-IQ4_NL	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ4_NL_XL	off	header	19.2/0.6	19.3/0.6	19.5/0.6	19.9/0.6	20.3/0.6	20.8/0.6	21.6/0.6	22.5/0.7	C	C
UD-IQ4_NL_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ4_XS	off	header	17.4/0.6	17.5/0.6	17.7/0.6	18.1/0.6	18.6/0.6	19.0/0.6	19.9/0.6	20.7/0.7	C	C
UD-IQ4_XS	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q2_K_XL	off	header	12.1/0.4	12.2/0.4	12.5/0.4	12.9/0.4	13.3/0.4	13.8/0.4	14.6/0.5	15.5/0.5	C	C
UD-Q2_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q3_K_M	off	header	16.4/0.4	16.5/0.4	16.7/0.4	17.1/0.5	17.6/0.5	18.0/0.5	18.9/0.5	19.7/0.6	C	C
UD-Q3_K_M	on	header	H	H	H	H	H	H	H	H	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.6 35B-A3B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q3_K_S	off	header	15.1/0.4	15.2/0.4	15.5/0.4	15.9/0.5	16.3/0.5	16.8/0.5	17.6/0.5	18.5/0.6	C	C
UD-Q3_K_S	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q3_K_XL	off	header	16.5/0.6	16.6/0.6	16.8/0.6	17.3/0.6	17.7/0.6	18.1/0.6	19.0/0.6	19.9/0.7	C	C
UD-Q3_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q4_K_M	off	header	21.8/0.6	21.9/0.6	22.1/0.6	22.5/0.6	23.0/0.6	23.4/0.6	24.3/0.6	25.1/0.7	C	C
UD-Q4_K_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q4_K_S	off	header	20.6/0.6	20.7/0.6	20.9/0.6	21.3/0.6	21.7/0.6	22.2/0.6	23.0/0.6	23.9/0.7	C	C
UD-Q4_K_S	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q4_K_XL	off	header	22.0/0.6	22.1/0.6	22.3/0.6	22.8/0.6	23.2/0.6	23.6/0.6	24.5/0.6	25.4/0.7	C	C
UD-Q4_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q5_K_M	off	header	26.1/0.6	26.2/0.6	26.4/0.6	26.9/0.6	27.3/0.6	27.7/0.6	28.6/0.6	29.5/0.7	C	C
UD-Q5_K_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q5_K_S	off	header	24.6/0.6	24.7/0.6	24.9/0.6	25.4/0.6	25.8/0.6	26.2/0.6	27.1/0.6	27.9/0.7	C	C
UD-Q5_K_S	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q5_K_XL	off	header	26.2/0.6	26.4/0.6	26.6/0.6	27.0/0.6	27.4/0.6	27.9/0.6	28.7/0.6	29.6/0.7	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.6 35B-A3B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q5_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q6_K	off	header	29.0/0.6	29.1/0.6	29.3/0.6	29.7/0.6	30.2/0.6	30.6/0.6	31.4/0.6	32.3/0.7	C	C
UD-Q6_K	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q6_K_XL	off	header	31.5/0.6	31.6/0.6	31.8/0.6	32.3/0.6	32.7/0.6	33.1/0.6	34.0/0.6	34.8/0.7	C	C
UD-Q6_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q8_K_XL	off	header	38.1/0.6	38.2/0.6	38.4/0.6	38.9/0.6	39.3/0.6	39.7/0.6	40.6/0.6	41.5/0.7	C	C
UD-Q8_K_XL	on	header	H	H	H	H	H	H	H	H	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Flint 40B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	79.2/0.6	79.3/0.6	79.6/0.6	80.2/0.7	80.7/0.7	81.3/0.7	82.5/0.7	83.6/0.8	C	C
BF16	on	header	H	H	H	H	H	H	H	H	C	C
Q8_0	off	file	42.2/0.3	42.3/0.3	42.6/0.4	43.2/0.4	43.8/0.4	44.3/0.4	45.5/0.4	46.6/0.5	C	C
Q8_0	on	file	H	H	H	H	H	H	H	H	C	C
UD-IQ1_M	off	header	11.4/0.1	11.6/0.1	11.8/0.2	12.4/0.2	13.0/0.2	13.6/0.2	14.7/0.2	15.9/0.3	C	C
UD-IQ1_M	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ2_XXS	off	header	11.9/0.1	12.1/0.1	12.4/0.2	12.9/0.2	13.5/0.2	14.1/0.2	15.2/0.2	16.4/0.3	C	C
UD-IQ2_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-IQ3_XXS	off	header	14.5/0.3	14.6/0.3	14.9/0.3	15.5/0.3	16.1/0.3	16.6/0.3	17.8/0.4	18.9/0.4	C	C
UD-IQ3_XXS	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q2_K_XL	off	header	13.6/0.2	13.7/0.2	14.0/0.2	14.6/0.2	15.2/0.2	15.7/0.2	16.9/0.3	18.0/0.3	C	C
UD-Q2_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q3_K_XL	off	header	18.4/0.3	18.5/0.3	18.8/0.4	19.4/0.4	19.9/0.4	20.5/0.4	21.7/0.4	22.8/0.5	C	C
UD-Q3_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q4_K_XL	off	header	24.9/0.3	25.1/0.3	25.4/0.4	25.9/0.4	26.5/0.4	27.1/0.4	28.2/0.4	29.4/0.5	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Flint 40B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q4_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q5_K_XL	off	header	29.8/0.3	29.9/0.3	30.2/0.4	30.8/0.4	31.4/0.4	31.9/0.4	33.1/0.4	34.2/0.5	C	C
UD-Q5_K_XL	on	header	H	H	H	H	H	H	H	H	C	C
UD-Q6_K_XL	off	file	36.2/0.3	36.3/0.3	36.6/0.4	37.2/0.4	37.8/0.4	38.3/0.4	39.5/0.4	40.6/0.5	C	C
UD-Q6_K_XL	on	file	H	H	H	H	H	H	H	H	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 26B-A4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
BF16	off	header	50.8/1.5	50.9/1.5	51.2/1.5	51.7/1.5	52.2/1.5	52.7/1.6	53.7/1.6	54.7/1.6	C	C
BF16	on	header	51.3/1.8	51.5/1.8	51.8/1.8	52.5/1.8	53.1/1.9	53.7/1.9	55.0/2.0	56.3/2.0	C	C
MXFP4_MOE	off	header	16.9/0.8	17.0/0.8	17.2/0.8	17.7/0.8	18.2/0.8	18.7/0.9	19.7/0.9	20.7/0.9	C	C
MXFP4_MOE	on	header	17.4/1.1	17.5/1.1	17.9/1.1	18.5/1.2	19.2/1.2	19.8/1.2	21.1/1.3	22.4/1.4	C	C
Q4_0	off	header	14.9/0.8	15.1/0.8	15.3/0.8	15.8/0.8	16.3/0.8	16.8/0.9	17.8/0.9	18.8/0.9	C	C
Q4_0	on	header	15.5/1.1	15.6/1.1	15.9/1.1	16.6/1.2	17.2/1.2	17.9/1.2	19.1/1.3	20.4/1.4	C	C
Q8_0	off	file	27.2/0.8	27.3/0.8	27.5/0.8	28.0/0.8	28.5/0.8	29.0/0.9	30.0/0.9	31.0/0.9	C	C
Q8_0	on	file	27.7/1.1	27.9/1.1	28.2/1.1	28.8/1.2	29.5/1.2	30.1/1.2	31.4/1.3	32.7/1.4	C	C
UD-IQ2_M	off	header	10.3/0.5	10.4/0.5	10.7/0.5	11.2/0.6	11.7/0.6	12.2/0.6	13.2/0.6	14.2/0.7	C	C
UD-IQ2_M	on	header	10.9/0.8	11.0/0.8	11.3/0.8	12.0/0.9	12.6/0.9	13.3/0.9	14.5/1.0	15.8/1.1	C	C
UD-IQ2_XXS	off	header	10.2/0.4	10.4/0.4	10.6/0.4	11.1/0.5	11.6/0.5	12.1/0.5	13.1/0.5	14.1/0.6	C	C
UD-IQ2_XXS	on	header	10.8/0.7	10.9/0.7	11.2/0.8	11.9/0.8	12.5/0.8	13.2/0.9	14.5/0.9	15.7/1.0	C	C
UD-IQ3_S	off	header	11.6/0.6	11.7/0.6	12.0/0.6	12.5/0.7	13.0/0.7	13.5/0.7	14.5/0.7	15.5/0.8	C	C
UD-IQ3_S	on	header	12.1/0.9	12.3/0.9	12.6/0.9	13.3/1.0	13.9/1.0	14.5/1.0	15.8/1.1	17.1/1.2	C	C
UD-IQ3_XXS	off	header	11.7/0.6	11.9/0.6	12.1/0.6	12.6/0.7	13.1/0.7	13.6/0.7	14.6/0.7	15.6/0.8	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 26B-A4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-IQ3_XXS	on	header	12.3/0.9	12.4/0.9	12.7/0.9	13.4/1.0	14.0/1.0	14.7/1.0	15.9/1.1	17.2/1.2	C	C
UD-IQ4_NL	off	header	13.9/0.8	14.0/0.8	14.3/0.8	14.8/0.8	15.3/0.8	15.8/0.9	16.8/0.9	17.8/0.9	C	C
UD-IQ4_NL	on	header	14.5/1.1	14.6/1.1	14.9/1.1	15.6/1.2	16.2/1.2	16.9/1.2	18.1/1.3	19.4/1.4	C	C
UD-IQ4_XS	off	header	13.9/0.8	14.0/0.8	14.3/0.8	14.8/0.8	15.3/0.8	15.8/0.9	16.8/0.9	17.8/0.9	C	C
UD-IQ4_XS	on	header	14.4/1.1	14.6/1.1	14.9/1.1	15.6/1.2	16.2/1.2	16.8/1.2	18.1/1.3	19.4/1.4	C	C
UD-Q2_K_XL	off	header	10.9/0.5	11.0/0.5	11.2/0.5	11.7/0.6	12.2/0.6	12.7/0.6	13.7/0.6	14.7/0.7	C	C
UD-Q2_K_XL	on	header	11.4/0.8	11.5/0.8	11.9/0.8	12.5/0.9	13.1/0.9	13.8/0.9	15.1/1.0	16.4/1.1	C	C
UD-Q3_K_M	off	header	13.0/0.6	13.2/0.6	13.4/0.6	13.9/0.7	14.4/0.7	14.9/0.7	15.9/0.7	16.9/0.8	C	C
UD-Q3_K_M	on	header	13.6/0.9	13.7/0.9	14.0/0.9	14.7/1.0	15.3/1.0	16.0/1.0	17.3/1.1	18.5/1.2	C	C
UD-Q3_K_XL	off	header	13.2/0.8	13.3/0.8	13.6/0.8	14.1/0.8	14.6/0.8	15.1/0.9	16.1/0.9	17.1/0.9	C	C
UD-Q3_K_XL	on	header	13.7/1.1	13.9/1.1	14.2/1.1	14.9/1.2	15.5/1.2	16.2/1.2	17.4/1.3	18.7/1.4	C	C
UD-Q4_K_M	off	header	17.3/0.8	17.4/0.8	17.6/0.8	18.1/0.8	18.6/0.8	19.1/0.9	20.1/0.9	21.1/0.9	C	C
UD-Q4_K_M	on	header	17.8/1.1	17.9/1.1	18.3/1.1	18.9/1.2	19.6/1.2	20.2/1.2	21.5/1.3	22.8/1.4	C	C
UD-Q4_K_S	off	header	16.8/0.8	16.9/0.8	17.2/0.8	17.7/0.8	18.2/0.8	18.7/0.9	19.7/0.9	20.7/0.9	C	C
UD-Q4_K_S	on	header	17.3/1.1	17.5/1.1	17.8/1.1	18.4/1.2	19.1/1.2	19.7/1.2	21.0/1.3	22.3/1.4	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Gemma 4 26B-A4B

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
UD-Q4_K_XL	off	header	17.3/0.8	17.4/0.8	17.7/0.8	18.2/0.8	18.7/0.8	19.2/0.9	20.2/0.9	21.2/0.9	C	C
UD-Q4_K_XL	on	header	17.8/1.1	18.0/1.1	18.3/1.1	19.0/1.2	19.6/1.2	20.3/1.2	21.5/1.3	22.8/1.4	C	C
UD-Q5_K_M	off	header	21.5/0.8	21.6/0.8	21.8/0.8	22.3/0.8	22.8/0.8	23.3/0.9	24.3/0.9	25.3/0.9	C	C
UD-Q5_K_M	on	header	22.0/1.1	22.1/1.1	22.5/1.1	23.1/1.2	23.8/1.2	24.4/1.2	25.7/1.3	27.0/1.4	C	C
UD-Q5_K_S	off	header	19.2/0.8	19.3/0.8	19.5/0.8	20.0/0.8	20.5/0.8	21.0/0.9	22.0/0.9	23.0/0.9	C	C
UD-Q5_K_S	on	header	19.7/1.1	19.8/1.1	20.2/1.1	20.8/1.2	21.5/1.2	22.1/1.2	23.4/1.3	24.7/1.4	C	C
UD-Q5_K_XL	off	header	21.5/0.8	21.7/0.8	21.9/0.8	22.4/0.8	22.9/0.8	23.4/0.9	24.4/0.9	25.4/0.9	C	C
UD-Q5_K_XL	on	header	22.1/1.1	22.2/1.1	22.5/1.1	23.2/1.2	23.8/1.2	24.5/1.2	25.7/1.3	27.0/1.4	C	C
UD-Q6_K	off	header	23.5/0.8	23.6/0.8	23.9/0.8	24.4/0.8	24.9/0.8	25.4/0.9	26.4/0.9	27.3/0.9	C	C
UD-Q6_K	on	header	24.0/1.1	24.2/1.1	24.5/1.1	25.1/1.2	25.8/1.2	26.4/1.2	27.7/1.3	29.0/1.4	C	C
UD-Q6_K_XL	off	header	23.6/0.8	23.7/0.8	24.0/0.8	24.5/0.8	25.0/0.8	25.5/0.9	26.5/0.9	27.5/0.9	C	C
UD-Q6_K_XL	on	header	24.1/1.1	24.3/1.1	24.6/1.1	25.3/1.2	25.9/1.2	26.5/1.2	27.8/1.3	29.1/1.4	C	C
UD-Q8_K_XL	off	header	27.9/0.8	28.1/0.8	28.3/0.8	28.8/0.8	29.3/0.8	29.8/0.9	30.8/0.9	31.8/0.9	C	C
UD-Q8_K_XL	on	header	28.5/1.1	28.6/1.1	29.0/1.1	29.6/1.2	30.2/1.2	30.9/1.2	32.2/1.3	33.4/1.4	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending

Memory matrix / Qwen3.6 35B-A3B coding

Metal / host GB by context window. Each column is a capacity estimate. The row identifies precision, speculative decoding and source type.

Quantization	MTP	Source	8K	16K	32K	64K	96K	128K	192K	256K	512K	1M
Q4_K_M	off	file	21.1/0.3	21.2/0.3	21.4/0.3	21.8/0.3	22.3/0.3	22.7/0.4	23.6/0.4	24.4/0.4	C	C
Q4_K_M	on	file	21.8/0.3	22.0/0.3	22.3/0.3	22.8/0.4	23.4/0.4	24.0/0.4	25.1/0.5	26.3/0.6	C	C

RAM-matrix-EN.csv · C: beyond native context H: no compatible head E: unavailable P: pending