STUDY · SPARKTEST

Models and hardware for SparkTest on premise

CONTENTS

Summary 3
- What we found 3
- Three configuration classes 5

1. What SparkTest asks of a model 6
- 1.1 Where the model is called 6
- 1.2 Prompt and answer sizes 7
- 1.3 Calls per test and per night 7
- 1.4 Latency and concurrency 8
- 1.5 Minimum model requirements 9

2. Selection criteria 9

3. The models 10
- 3.1 Node-scale candidates 10
- 3.2 Data-centre models 11
- 3.3 Licence conditions that need a decision 13
- 3.4 Notes on the landscape 13

4. The hardware 14
- 4.1 Class 1: compact unified-memory systems 14
- 4.2 Class 2: workstation GPUs 15
- 4.3 Class 3: data-centre accelerators and servers 15
- 4.4 What the hardware tables imply 17

5. Fit and speed 18
- 5.1 Memory per model (estimate) 18
- 5.2 Fit matrix (estimate) 19
- 5.3 Speed: estimates beside public measurements 21
- 5.4 Concurrency 22
- 5.5 Data-centre throughput (measured, no estimate made) 22
- 5.6 End-to-end agent measurements 23
- 5.7 Measured on SparkTest's own tasks 23

6. Configurations compared and recommended 24
- 6.1 Class 1: compact unified-memory system 27
- 6.2 Class 2: single workstation GPU 27
- 6.3 Class 3: multi-GPU server 28
- 6.4 Choice by size 28

7. Bringing your own model 29
- 7.1 Requirements 29
- 7.2 Connecting the endpoint 29
- 7.3 Serving software: compatibility on 16 and 17 November 2026 31
- 7.4 Checklist for the customer's IT team 32
- 7.5 Qualification with a test set 32

8. Risks and limits 33
- 8.1 Declared benchmarks and who measured them 33
- 8.2 Saturation and contamination 33
- 8.3 Quality loss from quantization 33
- 8.4 Licences 33
- 8.5 Prices and availability 33
- 8.6 Limits of the estimates and of the measurements 34
- 8.7 Italian 34

9. Method and reproducibility 34
- 9.1 Sources and labels 34
- 9.2 Formulas 35
- 9.3 Calibration values (computed by the calculator from `dati/measurements.json`) 35
- 9.4 Assumptions 36
- 9.5 Files and how to rerun 36

References 36
- SparkTest workload and serving software (note `fonti/workload-serving.md`) 36
- Models (note `fonti/models.md`) 38
- Hardware (note `fonti/hardware.md`) 42
- Inference measurements (note `fonti/measurements.md`) 45
- Benchmarks (note `fonti/benchmarks.md`) 47
- Method references added by the author 48

Verification 48

A study of open-weight language models, inference hardware and customer-managed model endpoints

SparkTech, Turin. Date of the study: 16 November 2026; revised on 17 November 2026 after an independent check of its figures, then updated the same day for the product changes of 17 November 2026 and with SparkTech's first measurement of SparkTest's own tasks on open-weight models (section 5.7; see "Verification" at the end). All data as of 16 November 2026 unless dated otherwise; every web source was accessed on that date unless its reference says otherwise.

How to read this study. Bracketed codes such as [models-12] or [hardware-9] point to the numbered references at the end. Every figure has a source, and the tables and the summary say what kind of figure it is:

- **declared**: stated by the producer (model developer, hardware vendor, software project);
- **measured**: measured by someone other than this study; the text says who: an independent party (analyst, benchmark consortium, press, community), a runtime developer, or an MLPerf submitter (results peer-reviewed by MLCommons);
- **observed**: a price read on a dated public page;
- **estimate**: computed by this study's calculator (`calcolatore.mjs`), never a measurement;
- **measured by SparkTech**: measured by SparkTech on SparkTest's own tasks, on a test server, for the companion document summarised in section 5.7; never the speed of a node.

Figures about SparkTest itself come from its source code as of 16 November 2026, updated for the changes of 17 November 2026 (internal references [workload-serving-1] to [workload-serving-10] and [workload-serving-67], which have no public URL); the calls and tokens per night derived from them are estimates. Where a figure was searched for and not found, the text says "not found".

Terms. A *token* is a piece of text, about four characters of English. *Prefill* is the reading of the prompt and *decode* the writing of the answer, both in tokens per second (tok/s). The *KV cache* is the memory each open request needs; a *32k session* is one request of up to 32,768 tokens (context lengths use k = 1,024 tokens). A *mixture-of-experts (MoE)* model uses only part of its parameters, the *active parameters*, for each token; a *dense* model uses all of them. *BF16, 8-bit, 4-bit* say how many bits store each weight: fewer bits need less memory and run faster, with a possible loss of quality (section 8.3). A *burst* is all the concurrent requests of a size arriving together. The *target* is the time this study aims for; the *hard limit* is the time after which the agent stops a call.

Summary

What we found

01 **SparkTest asks little of a model in volume and a lot in latency.** A run whose test plan still holds makes no model call. The model works when a test is written or edited (8 to 16 calls of about 4,300 to 4,900 prompt tokens and 60 to 120 answer tokens), when a run fails (1 to 2 repair calls of about 3,200 to 3,700 prompt tokens and 440 to 720 answer tokens, then one diagnosis if the failure persists) and on request [workload-serving-1][workload-serving-4][workload-serving-10]. A failed test is not codified again: since 17 November 2026 its next run executes the same plan and repairs it under the agent's scenario check [workload-serving-67]. A typical night with 400 maintained tests needs about 54 to 106 calls; a release night with 1,500 tests about 590 to 1,160 calls and 2.2 to 5.1 million prompt tokens (estimate, section 1.3) [workload-serving-10]. What binds is the time of each call while 2, 4 or 8 parallel runs (sizes S, M, L) and one API analysis are in flight: the agent stops a codify or repair call at 60 s and a diagnosis at 30 s [workload-serving-1][workload-serving-4][workload-serving-9].

02 **One class of model fits a compact system at these latencies.** MoE models with about 3B active parameters and hybrid attention (Qwen3.6-35B-A3B, NVIDIA Nemotron 3.5 Lightning 30B-A3B, poolside Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b) need 12 to 20 GB of weights at 4 bits and 0.2 to 1.4 GB of KV cache per 32k session (estimate, section 5.1). Dense 24B to 32B models decode at 6 to 12 tok/s on compact unified-memory systems (measured by a runtime developer and by the community [measurements-11][measurements-20][measurements-21]) and belong on a workstation GPU. Models of 117B to 125B parameters held in memory (gpt-oss-120b, Mistral Small 4, Nemotron 3.5 Super VL, Qwen3.5-122B-A10B) fit 128 GB at 4 bits, but none meets the targets on a compact system: gpt-oss-120b and Mistral Small 4 stay within the agent's hard limits up to size M on GB10, and on Ryzen AI Max+ 395 only gpt-oss-120b does, up to size S; a 96 GB workstation GPU brings them within the targets (estimate, section 6). Data-centre models (180B to 2.8T parameters) need a multi-GPU server.

03 **Licences are the first filter.** Apache-2.0, MIT and OpenMDW-1.1 cover most node-scale candidates. Llama 4 grants no rights on its multimodal models to EU-domiciled companies; ERNIE 5.5 Flash Base excludes the EU; MiniMax M2.7 is non-commercial; Mistral Medium 3.5 and Devstral 2 give no rights above USD 20 million of monthly company revenue; Qwen3.8-Flash-Next needs a separate licence for "AI Work Assistant" products [models-44][models-45][models-162][models-146][models-59][models-61][models-22]. Who is the licensee when the model runs on a node in the customer's network is a legal question this study does not answer.

04 **No public evidence measures SparkTest's task.** On the Artificial Analysis Intelligence Index v5.0.1 the open-weight model ranked first scores 46 against 65 for the proprietary model ranked first; node-scale models have few scores (Qwen3.8-27B: 20; gpt-oss-120b and Gemma 4 31B: 4) (independent) [benchmarks-1][benchmarks-3][benchmarks-4][benchmarks-5]. This study found no independent leaderboard of JSON-schema adherence and no Italian-specific evaluation of 2026 open-weight models. For the same model, vendor and independent figures differ by up to 36 points [benchmarks-49][benchmarks-18][benchmarks-23]. A model must be qualified on SparkTest's own test set before use (section 7.5); SparkTech's first measurement of that kind is point 7.

05 **Bandwidth bounds generation; compute bounds prompt reading.** The two compact platforms compared have 256 GB/s (derived from the declared memory bus) and 273 GB/s (declared) and generate at similar speeds, but read prompts at different speeds: gpt-oss-120b prefill at 32k depth is 1,567 tok/s on DGX Spark (measured by the runtime developer) against 234 to 653 tok/s on Ryzen AI Max+ 395 depending on the software backend (community) [measurements-2][measurements-20]; in MLPerf Inference v6.1 Edge Agentic an agent turn took 3.81 s against 7.06 s (measured by the submitter, peer-reviewed) [measurements-26]. SparkTest prompts are long and answers short, so prompt reading matters.

06 **Prices are volatile.** DGX Spark list price: USD 3,999 (October 2025), 4,699 (February 2026), 6,950 (October 2026). A GeForce RTX 5090 sold from EUR 7,399.99 in Germany on 16 November 2026 against an EU list price of EUR 2,229 (observed) [hardware-3][hardware-4][hardware-5][hardware-44][hardware-45]. Every price here is a dated indication.

07 **Measured on SparkTest's own tasks (measured by SparkTech).** On a test server without a GPU, Qwen3.6-35B-A3B (Q4_K_M) and gpt-oss-20b (MXFP4), served by llama.cpp, answered the eight tasks of SparkTest's qualification kit in English and Italian, with the agent's own prompts and checks. gpt-oss-20b passed all 23 required checks, Qwen3.6 22 (one plan rejected by the agent's validator); every answer constrained by a schema passed the schema check; both Italian diagnoses cited the evidence. Both models changed an expected value when a repair prompt invited it, and the agent's scenario check rejected both repairs, so the regression stayed visible. The times (median 101 and 108 s per call, 1 call in 8 within the agent's limits) describe a 2016 server processor, not a node: prompt reading, which a CPU lacks, took most of each call (section 5.7).

Three configuration classes

Sizes: S = 100 maintained tests and 2 parallel runs, M = 400 and 4, L = 1,500 and 8 [workload-serving-9]. The verdicts below are estimates (sections 5 and 6) for a burst in which all concurrent calls of the size arrive together, with the concurrent 32k sessions in memory:

- **within targets**: the burst meets this study's targets (codify round 15 s, repair 30 s, diagnosis 30 s);
- **within hard limits**: the burst meets the agent's hard limits (60 s, 60 s, 30 s); SparkTest works, more slowly (section 6);
- **usable up to**: the largest size within the hard limits both with typical prompts (codify round of 4,941 tokens) and with application documentation at the 24,000-character cap (codify round of 11,093 tokens).

CLASS	HARDWARE COMPARED	MEMORY; BANDWIDTH	MODELS THAT FIT THE CLASS (ESTIMATE)	SIZES (ESTIMATE)	INDICATIVE PRICE, 16 NOV 2026 (OBSERVED)	DOES NOT DO
1. Compact unified-memory system	NVIDIA GB10 systems (DGX Spark and partner boxes); AMD Ryzen AI Max+ 395 and PRO 495 systems	64 to 192 GB; 256 GB/s (derived) or 273 GB/s (declared)	about 3B-active MoE at 4 bits: Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b	GB10 with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1: within targets up to M with typical prompts, S with documentation at the cap; usable up to L. GB10 with Gemma 4 26B-A4B or gpt-oss-20b: targets up to S, usable up to L. Ryzen AI Max+ 395: no size within targets; usable up to M (up to S with Gemma 4 26B-A4B or gpt-oss-20b)	from EUR 3,899.96 (Ryzen AI Max+ 395, 128 GB) and EUR 5,999 (GB10, 128 GB), street DE [hardware-22] [hardware-9]	size L within the targets; dense 24B to 32B models at the targets; 117B to 125B MoE within the targets (within the hard limits: up to M on GB10; on Ryzen AI Max+ 395 only gpt-oss-120b, up to S); data-centre models
2. Single workstation GPU	RTX PRO 6000 (96 GB), RTX PRO 5000 (48 or 72 GB), 32 GB cards (RTX PRO 4500, Radeon AI PRO R9700, Arc Pro B70; GeForce RTX 5090 with the licence restriction of section 4.2)	32 to 96 GB per card; 608 to 1,792 GB/s (declared)	96 GB: every node-scale candidate; 48 to 72 GB: 3B-active MoE at 4 to 8 bits, dense 27B to 31B at 4 to 8 bits; 32 GB: 3B-active MoE at 4 bits	96 GB: within targets up to L for 3B-active MoE, gpt-oss-120b and Mistral Small 4 (the last two: M with documentation at the cap); dense 27B at 8 bits: M. 48 to 72 GB: L for 3B-active MoE (M with documentation at the cap). 32 GB: M for Qwen3.6-35B-A3B, Nemotron 3.5 Lightning and gpt-oss-20b (S with documentation at the cap), usable up to L; RTX 5090: L	card from EUR 1,659 (Arc Pro B70) to EUR 16,489.98 (RTX PRO 6000 Max-Q), street DE; host computer extra [hardware-61] [hardware-48]	117B to 125B MoE below 96 GB; data-centre models; several cards carry datacenter restrictions (section 4.2)
3. Multi-GPU server	4 x RTX PRO 6000 Server Edition; 8 x H100, H200 or B200 (HGX); 8 x Instinct MI355X	384 GB to 2.3 TB	data-centre models: Qwen3.5-397B-A17B, GLM-5.3-Flash, DeepSeek-V4-Flash, Mistral Large 3, Nemotron 3 Ultra; GLM-5.3 and Kimi K3.1 Pro under custom licences	not limited by SparkTest's load (measured throughput, section 5.5)	USD 195,000 (used 8 x H100 node) to USD 470,000 (new 8 x B200 node), US, October 2026 [hardware-87][hardware-88]	it is not an appliance; cost and power (up to about 14 kW for a DGX B200 or DGX B300 system, declared) are of another order [hardware-68][hardware-69]

Bringing your own model. A customer that already runs a model can connect it through an OpenAI-compatible endpoint (`LLM_URL`). It needs server-side JSON-schema enforcement, at least 32k tokens of context, reasoning switched off, as many concurrent requests as the size's parallel runs plus one, and per-call times within the targets. Until 16 November 2026 the agent's request was accepted by llama.cpp server and Ollama, while vLLM 0.33 and SGLang 0.5.23 rejected it because it carried no schema name. Since 17 November 2026 the agent names its schema and also asks for no reasoning through `chat_template_kwargs` and `reasoning_effort`, the fields these servers read; on the evidence of their source code vLLM and SGLang now accept the request, but no run against them has been made yet [workload-serving-16][workload-serving-24][workload-serving-31][workload-serving-40][workload-serving-67]. Section 7 lists the requirements, the changes still open and a qualification procedure.

What this means for your IT team

- If SparkTech supplies the node, SparkTech sizes each SparkNode on your volume with the method of section 9 and qualifies it on your own tests before use (section 7.5).
- If you bring your own model, check your endpoint against section 7.1 and the checklist of section 7.4, then run SparkTest's qualification kit on it (sections 5.7 and 7.5). Since 17 November 2026 the agent's request carries the schema name that vLLM, SGLang and NVIDIA NIM require; OpenVINO Model Server and TGI still need the changes listed in section 7.3.
- Plan for 3, 5 or 9 requests at once (sizes S, M, L), each with a prompt of 4,000 to 11,000 tokens. Nightly volume is small; the time of each call is what counts.
- Within the targets (estimate): a GB10 system up to size M with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1, a 32 GB workstation card up to M, an RTX PRO 5000 or RTX PRO 6000 up to L. Where application documentation reaches the cap, GB10 and the 32 GB cards drop to S and the RTX PRO 5000 to M. Confirm on site with section 7.5.
- Settle who holds the model licence on your premises before purchase (section 8.4).

1. What SparkTest asks of a model

SparkTest turns a test written in plain language into a JSON test plan, runs the plan in a browser, and repairs it when the application changes. This section describes the model load as read from the code; it does not depend on any model. Detailed figures and method: companion note `fonti/workload-serving.md` and data `dati/workload.json`.

1.1 Where the model is called

OPERATION	WHEN IT RUNS	ANSWER	SCHEMA ENFORCED BY THE SERVER	TIMEOUT PER CALL	SOURCES
Codify round	new or edited test, or a stored script without a plan; since 17 November 2026 never after a run that ended "failed" (the plan runs again and is repaired)	JSON: up to 5 next steps and a note	yes	60 s; whole exploration 4 min, at most 15 rounds	[workload-serving-1] [workload-serving-4] [workload-serving-67]
One-shot codify	browser unavailable or exploration budget used up	complete plan JSON (1 to 60 steps)	yes	60 s	[workload-serving-1] [workload-serving-2]
Repair	a run failed for a reason other than infrastructure	complete corrected plan JSON	yes	60 s; up to 2 attempts per run	[workload-serving-1] [workload-serving-4]
Diagnosis	a run ends "failed"	Markdown, 60 to 120 words, in the test's language	no	30 s	[workload-serving-1] [workload-serving-4]
API insights	on request, per application	JSON, at most 6 tips	yes	150 s	[workload-serving-1] [workload-serving-6]
Plain execution	nightly suite and on demand	none	none	none	[workload-serving-4] [workload-serving-6]

Every plan is validated by the agent even when the server enforced the schema. A repair that removes assertions or too much of the scenario is rejected, and since 17 November 2026 so is a repair that changes an expected value (a text or address to find, a minimum count, the kind of a presence check) or a value the test types (a field, a filter, the row it opens): selectors, labels and waits may change, what the test verifies may not. The run then fails as a candidate regression [workload-serving-1][workload-serving-2][workload-serving-67]. In the version examined, the customer endpoint serves the agent's operations above; the console's test suggestions and test chat use a separate model connection that `LLM_URL` does not configure, with prompts of up to about 26,000 tokens [workload-serving-5][workload-serving-10]. An on-premise installation must state where these two features get their model (the same local endpoint, another local model, or switched off), so that the data-perimeter review of section 7.1 covers them.

1.2 Prompt and answer sizes

Characters were measured by running the agent's own prompt builders on synthetic inputs (no model was called). Tokens: characters divided by 4.0 for English, OpenAI's rule of thumb [workload-serving-11], and by 3.5 for Italian, an assumption of the workload study [workload-serving-10].

OPERATION	CASE	PROMPT TOKENS EN / IT	ANSWER TOKENS EN / IT	SOURCES
Codify round	typical 20-step test, average of 11 rounds	4,324 / 4,941	62 / 70 (compact JSON)	[workload-serving-10]
Codify round	application documentation at the 24,000-character cap	9,707 / 11,093	62 / 70	[workload-serving-10]
Repair	typical 20-step plan	3,233 / 3,695	436 to 631 / 498 to 721	[workload-serving-10]
Repair	25 steps (the maintained-test limit)	3,384 / 3,867	540 to 781 / 617 to 893	[workload-serving-9][workload-serving-10]
Diagnosis	typical	1,053 / 1,203	about 210 / 240 (estimate)	[workload-serving-10]
API insights	typical (25 endpoints)	4,700 / 5,372	366 / 418	[workload-serving-10]
API insights	every field at its cap	17,076 / 19,515	2,155 / 2,463	[workload-serving-10]

The page snapshot (up to 60 interactive elements and 3,000 characters of visible text), the inlined documentation and, in a repair, the plan itself dominate the prompts [workload-serving-1][workload-serving-3]. The answer budget set by default in the agent (2,048 tokens) is too small for six long API tips (2,463 tokens); the workload study recommends 4,096 [workload-serving-1][workload-serving-10].

1.3 Calls per test and per night

EVENT	MODEL CALLS	SOURCES
Run with a plan that holds	0	[workload-serving-4]
Codify a 20-step test	8 to 16	[workload-serving-4] [workload-serving-10]
Repair a failing run	1 to 2	[workload-serving-4]
Real defect (repairs do not make it pass)	1 to 2 repairs and 1 diagnosis; at the next run the same plan runs again (until 16 November 2026 the next run codified the test again, 8 to 16 calls)	[workload-serving-4] [workload-serving-67]

Night scenarios. The shares of tests that break, fail without a change, fail for a real defect or are edited are assumptions of the workload study (UI breakage 1%, 3%, 10%; flaky 1.5%, an order of magnitude reported by Google for all its test types in 2016; real defects 0.2%, 0.5%, 2%; edited tests 0.5%, 1%, 3%) [workload-serving-9][workload-serving-63][workload-serving-10]. A real defect costs its repairs and a diagnosis, with no new codification (the rule of 17 November 2026; the figures computed on 16 November with a new codification are in "Verification") [workload-serving-67]. Results are bounds, not forecasts (estimate):

SCENARIO	TESTS	CALLS	PROMPT TOKENS	ANSWER TOKENS	SOURCES
Typical	100	14 to 27	51,283 to 116,616	2,775 to 8,450	[workload-serving-10]
Typical	400	54 to 106	205,130 to 466,463	11,098 to 33,801	[workload-serving-10]
Typical	1,500	203 to 398	769,238 to 1,749,236	41,618 to 126,754	[workload-serving-10]
Release night	100	40 to 77	149,524 to 339,363	7,775 to 23,309	[workload-serving-10]
Release night	400	158 to 308	598,096 to 1,357,453	31,101 to 93,237	[workload-serving-10]
Release night	1,500	593 to 1,155	2,242,858 to 5,090,447	116,629 to 349,637	[workload-serving-10]

Spread over a 6-hour night, even the release night with 1,500 tests averages 104 to 236 prompt tokens per second [workload-serving-10]. Volume is not the constraint; the time of each call is.

1.4 Latency and concurrency

One node runs up to 2, 4 or 8 test runs at a time (sizes S, M, L) and one API analysis beside them, so the endpoint must serve 3, 5 or 9 concurrent requests [workload-serving-4][workload-serving-9]. The time of a call is roughly prompt tokens / prefill speed + answer tokens / decode speed + waiting time in the server's queue; the agent's timer includes the wait [workload-serving-10].

OPERATION (ITALIAN TOKEN COUNTS)	HARD LIMIT IN THE CODE	TARGET USED IN THIS STUDY	DECODE NEEDED PER REQUEST AT A PREFILL OF 500 / 2,000 TOK/S	SOURCES
Codify round, typical (4,941 in, 114 out)	60 s	15 s (11 rounds and browser time must fit 4 minutes)	22.3 / 9.1 tok/s	[workload-serving-10]
Codify round, documentation at cap (11,093 in, 114 out)	60 s	15 s	not reachable / 12.1 tok/s	[workload-serving-10]
Repair, 20 steps (3,695 in, 721 out)	60 s	30 s (the run waits for it)	31.9 / 25.6 tok/s	[workload-serving-10]
Diagnosis, typical (1,203 in, 240 out)	30 s	30 s	8.7 / 8.2 tok/s	[workload-serving-10]
API insights, typical (5,372 in, 418 out)	150 s	150 s	3.0 / 2.8 tok/s	[workload-serving-10]

The codify rows use 114 answer tokens: the 400 characters a round's answer may take when the server allows whitespace, a conservative case against the 62 to 70 tokens of compact JSON in section 1.2 [workload-serving-10].

In short: about 500 to 2,000 tok/s of prompt reading and 25 to 32 tok/s of generation per request, with 3 to 9 requests in flight. A server that queues requests (Ollama serves one request at a time per model by default) adds the waiting time inside the same limits [workload-serving-36]. After a timeout the agent retries once with a shorter prompt [workload-serving-1].

1.5 Minimum model requirements

REQUIREMENT	VALUE	SOURCES
Structured output	JSON that follows a JSON Schema, enforced by the server (`response_format` of type `json_schema`)	[workload-serving-1] [workload-serving-2]
Schema keywords	`type`, `properties`, `required`, `additionalProperties: false`, `enum`, `items`, `minItems` / `maxItems`, anchored `pattern`, integer `minimum` / `maximum`, `maxLength`	[workload-serving-2]
Context	32k tokens minimum (the theoretical codify maximum, 33,741 tokens with the answer, exceeds 32,768); 64k recommended	[workload-serving-10]
Answer budget	4,096 tokens recommended (default 2,048)	[workload-serving-1] [workload-serving-10]
Languages	instructions in English; descriptions, documentation and page text in Italian or English; diagnosis and tips written in the test's language	[workload-serving-1] [workload-serving-5]
Reasoning	off: the agent expects direct answers; reasoning tokens would consume the answer budget and the timeouts	[workload-serving-1]
Determinism	temperature 0.000001, seed 0	[workload-serving-1]

2. Selection criteria

The 120 open-weight models reviewed (companion data `dati/models.json`) were filtered in this order:

01 **Licence.** The model must be usable commercially by an EU company on a node in the customer's network. Licences that exclude the EU, forbid commercial or production use, or carry business-model clauses are excluded or flagged for legal review (table 3.3).

02 **Memory.** Weights at a chosen precision plus nine 32k sessions (size L) must fit in the memory left after a reserve (section 9). A model that holds fewer sessions serves smaller sizes only.

03 **Speed.** Per-call times within the targets of section 1.4 for the size, first as an estimate (section 5), then measured on the customer's hardware (section 7.5).

04 **Structured output and tool use.** JSON-schema enforcement is a feature of the inference engine, not of the model: vLLM enforces schemas through its xgrammar or guidance backends, llama.cpp through grammars [models-175][models-176]. Model cards almost always declare tool calling; explicit structured-output claims are rare (gpt-oss "Structured Outputs", Mistral "JSON output", Nemotron "structured outputs" as a training target) [models-37][models-56][models-118].

05 **Italian.** Declared in the model card, or not. Qwen and Gemma cards claim 140 to 200+ languages without naming Italian; GLM lists English and Chinese only [models-3][models-71][models-79]. Declared is not measured (section 8.7).

06 **Context.** At least 32k tokens native.

07 **Reasoning off.** SparkTest runs with reasoning switched off, so a model's score in non-reasoning mode is the relevant one where published: Qwen3.8-27B scores 20 on the Artificial Analysis index at the "xhigh" effort setting and 10 without reasoning (independent) [models-16].

08 **Engine support.** Hybrid attention (Gated DeltaNet, Mamba-2, sliding windows), multi-head latent attention (MLA) and sparse indexers must be supported by the chosen engine version, together with schema-constrained decoding for the model's tokenizer [models-175][models-176].

09 **Independent evidence**, where it exists, with its harness noted (section 8.1).

For your IT team: apply the same filters, in the same order, to any model you plan to bring (section 7).

3. The models

Shortlist: 19 node-scale models (one device or one compact system) and 9 data-centre models (multi-GPU servers; relevant for customers that run their own endpoint). Values are declared by the developer unless marked; "Index" gives the Artificial Analysis Intelligence Index v5.0.1 (AA, independent) and the Epoch Capabilities Index with its 95% interval (ECI, independent) where they exist.

3.1 Node-scale candidates

MODEL	DEVELOPER, RELEASE	TOTAL / ACTIVE PARAMETERS	LAYERS THAT KEEP A GROWING KV CACHE	NATIVE CONTEXT	LICENCE	ITALIAN IN THE CARD	TOOLS / JSON (DECLARED)	INDEX (AA; ECI)	SOURCES
Qwen3.6-35B-A3B	Qwen (Alibaba), 2026-04	35B / 3B MoE	10 of 40 (others Gated DeltaNet)	256k (262,144)	Apache-2.0	not listed	tools	AA no score shown; ECI 144.0 (142.1 to 145.9)	[models-12][models-14] [benchmarks-15]
Qwen3.8-27B	Qwen, 2026-08	27B dense	16 of 64	256k (262,144)	Apache-2.0	not listed	tools	AA 20 (xhigh), 10 (no reasoning); ECI 152.6 (150.6 to 154.8)	[models-15][models-16] [benchmarks-15]
Qwen3.5-122B-A10B	Qwen, 2026-02	122B / 10B MoE	12 of 48	256k (262,144)	Apache-2.0	not listed (201 languages claimed)	tools	not found	[models-10][models-3]
Qwen3-Coder-Next	Qwen, 2026-01	80B / 3B MoE	12 of 48	256k (262,144)	Apache-2.0	not listed	tools	not found	[models-6]
gpt-oss-20b	OpenAI, 2025-08	21B / 3.6B MoE	12 of 24 (12 sliding, window 128)	128k (131,072)	Apache-2.0	not stated	tools; Structured Outputs	AA no score shown; ECI 135.0 (132.3 to 137.5)	[models-36][models-37] [models-38][benchmarks-15]
gpt-oss-120b	OpenAI, 2025-08	117B / 5.1B MoE	18 of 36 (18 sliding, window 128)	128k (131,072)	Apache-2.0	not stated	tools; Structured Outputs	AA 4 (high); ECI 140.2 (138.2 to 142.1)	[models-39][models-37] [models-40][benchmarks-15]
Gemma 4 26B-A4B	Google, 2026-04	25.2B / 3.8B MoE	5 of 30 (25 sliding, window 1,024)	256k	Apache-2.0	not listed (140+ languages in pre-training, 35+ supported)	tools	AA no score shown; ECI 141.6 (139.1 to 144.1)	[models-74][models-71] [models-75][benchmarks-15]
Gemma 4 31B	Google, 2026-04	30.7B dense	10 of 60 (50 sliding, window 1,024)	256k	Apache-2.0	not listed (140+ languages in pre-training, 35+ supported)	tools	AA 4 (reasoning); ECI 144.3 (141.7 to 146.8)	[models-72][models-73] [benchmarks-15]
Muse Glimmer 30B	Meta, 2026-08	29.6B dense (incl. vision encoder)	13 of 52 (39 sliding, window 2,048)	128k (131,072) or more	Apache-2.0	not listed (100+ languages)	tools with precise schemas	not found	[models-48]
Nemotron 3.5 Lightning 30B-A3B	NVIDIA, 2026-08	30B / 3B MoE	6 attention layers (23 Mamba-2)	up to 1M	OpenMDW-1.1	listed	tools; structured outputs in training data	not found in v5.0.1	[models-118][models-119]

MODEL	DEVELOPER, RELEASE	TOTAL / ACTIVE PARAMETERS	LAYERS THAT KEEP A GROWING KV CACHE	NATIVE CONTEXT	LICENCE	ITALIAN IN THE CARD	TOOLS / JSON (DECLARED)	INDEX (AA; ECI)	SOURCES
Nemotron 3.5 Bolt 74B-A9B	NVIDIA, 2026-10	74B / 9B MoE	8 attention layers (40 Mamba-2)	up to 256k	OpenMDW-1.1	listed	tools	not found	[models-121][models-119]
Nemotron 3.5 Super VL 120B-A12B	NVIDIA, 2026-10	120B / 12B MoE	8 attention layers (40 Mamba-2)	up to 1M	OpenMDW-1.1	listed	tools	AA 14	[models-122][models-17]
Mistral Small 4 119B	Mistral AI, 2026-03	119B / 6.5B MoE	36 MLA layers	256k	Apache-2.0	listed	tools; JSON output	AA 4 (reasoning)	[models-56][models-57]
Mistral Small 3.2 24B	Mistral AI, 2025-06	24B dense	40 of 40	128k (131,072)	Apache-2.0	listed	tools	not found (BFCL V4: 37.15%, hosted version)	[models-52][models-2]
Ministral 3 14B	Mistral AI, 2025-12	14B dense	40 of 40	256k (262,144)	Apache-2.0	listed	tools; JSON output	not found	[models-55]
Granite 4.2 30B	IBM, 2026-08	29.3B dense	64 of 64	128k (512k extension)	Apache-2.0	listed (12 tested languages)	tools	not found	[models-113]
Laguna XS 2.1	poolside, 2026-06	33B / 3B MoE	10 of 40 (30 sliding, window 512)	256k (262,144)	OpenMDW-1.1	not listed	tools	not found	[models-124][models-119]
GLM-4.7-Flash	Z.ai, 2026-01	31B / 3B MoE (declared as "30B-A3B")	47 MLA layers	198k (202,752)	MIT	not supported (English, Chinese)	tools	not found	[models-79]
EuroLLM-22B	UTTER project, 2025-12	22.6B dense	54 of 54	32k (32,768)	Apache-2.0	listed (all EU official languages)	not stated	not found	[models-132]

3.2 Data-centre models

MODEL	DEVELOPER, RELEASE	TOTAL / ACTIVE PARAMETERS	NATIVE CONTEXT	LICENCE	ITALIAN IN THE CARD	INDEX (AA; ECI)	SOURCES
Qwen3.8-Flash-Next	Qwen, 2026-08	125B + 51B n-gram embedding + 4B MTP (180B stored) / 6B	256k (262,144)	Qwen Community License 1.0	not listed	AA 29	[models-21][models-22][models-24]
GLM-5.3-Flash	Z.ai, 2026-08	320B / 18B	1M	MIT	not supported	AA 28 (model page); ECI 153.8 (152.0 to 155.8)	[models-87][models-88][benchmarks-15]
DeepSeek-V4-Flash-0731	DeepSeek, 2026-07	284B (304B stored) / 13B	1M	MIT	not listed	AA not found; ECI 154.4 (152.6 to 156.4)	[models-32][benchmarks-15]
Qwen3.5-397B-A17B	Qwen, 2026-02	397B / 17B	256k (262,144)	Apache-2.0	not listed	AA not found in v5.0.1; ECI 147.9 (145.8 to 150.1)	[models-20][benchmarks-15]
Nemotron 3 Ultra 550B-A55B	NVIDIA, 2026-06	550B / 55B	up to 1M	OpenMDW-1.1	listed	AA not found; ECI 147.0 (144.8 to 149.2)	[models-123][benchmarks-15]
Mistral Large 3 675B	Mistral AI, 2025-12	675B / 41B	256k	Apache-2.0	listed	not found	[models-64]

MODEL	DEVELOPER, RELEASE	TOTAL / ACTIVE PARAMETERS	NATIVE CONTEXT	LICENCE	ITALIAN IN THE CARD	INDEX (AA; ECI)	SOURCES
DeepSeek-V4.1-Flash-1017	DeepSeek, 2026-10	552B backbone + 196B memory module (763B stored) / 16B in decode	1M	MIT	not listed	AA 33 (max), 19 (no reasoning); ECI 159.4 (157.0 to 162.0)	[models-34][models-35][benchmarks-3] [benchmarks-11][benchmarks-15]
GLM-5.3	Z.ai, 2026-08	753B stored / 40B (the card states the same base model as GLM-5.2 and no count; 744B total and 40B active are declared for GLM-5)	1M	GLM-5.3 License	not supported	AA 32; ECI 157.1 (154.9 to 159.5)	[models-85][models-82][models-86] [benchmarks-15]
Kimi K3.1 Pro	Moonshot AI, 2026-11	2.8T / 104B	1M	Kimi K3.1-Pro License	not listed	AA 46; ECI 161.4 (159.0 to 164.0)	[models-100][models-101][models-17] [benchmarks-15]

For reference, the proprietary model ranked first on each board scores 65 (AA) and 173.4 (ECI) [benchmarks-4][benchmarks-15].

3.3 Licence conditions that need a decision

Licence texts as read on 16 November 2026; summary of the companion note, not legal advice.

LICENCE	MODELS	CONDITION AS WRITTEN	EFFECT FOR AN ON-PREMISE PRODUCT	SOURCES
Apache-2.0, MIT, OpenMDW-1.1	most node-scale candidates	permissive; OpenMDW-1.1 adds a patent-retaliation termination clause and no obligation on outputs	no business-model restriction found	[models-119]
NVIDIA Nemotron Open Model License	Nemotron Nano 9B v2, Nemotron 3 Nano and Super	commercial use, no revenue threshold; attribution; indemnification of NVIDIA, including for outputs	usable; indemnity clause to review	[models-115]
Qwen Community License 1.0	Qwen3.8-Flash-Next	separate licence, regardless of revenue, for "AI Work Assistant" (AI-assisted coding or office productivity) and Model-as-a-Service businesses	a product that turns test descriptions into executable test plans could be read as AI-assisted coding: legal opinion needed	[models-22]
Modified MIT (Mistral AI)	Mistral Medium 3.5, Devstral 2	no rights if the licensee's company has more than USD 20 million of monthly revenue	if the customer is the licensee, large customers are excluded	[models-59] [models-61]
Step 5 Preview licence	Step-5-Preview	separate licence for makers of hardware devices that offer the model as a device feature	an appliance offer may fall under it	[models-153]
Muse Model License	Muse Spark Open	"Built with Muse Spark Open" in every product that uses it	display obligation in SparkTest	[models-51]
Llama 4 Community License	Llama 4 Scout, Maverick	no rights for EU-domiciled individuals or companies on multimodal Llama 4 models; the clause sits in the Acceptable Use Policy that the licence incorporates and does not apply to end users of a product that includes the model	excluded for an Italian company that builds the model into its product	[models-44] [models-45]
ERNIE 5.5 Flash Base licence	ERNIE-5.5-Flash-Base	territory excludes the EU, the UK and South Korea	excluded	[models-162]
MiniMax M2.7 Non-Commercial	MiniMax-M2.7	commercial use only with prior written authorization	excluded	[models-146]
Mistral AI Non-Production License	Codestral-22B-v0.1	non-production use only	excluded	[models-63]

3.4 Notes on the landscape

- DeepSeek has no current release below 284B total parameters (V4-Flash) [models-30]. Meta's 2026 open models are published under the Muse name; no Llama release after Llama 4 was found [models-48][models-51]. Qwen4 has a hosted preview since 2026-11-11 with no weights, card or licence published [models-177][models-178].
- Most 2026 releases keep a growing KV cache in only some of their layers (Gated DeltaNet, Mamba-2, sliding windows), which is why a 32k session costs 0.2 GB on Nemotron 3.5 Lightning against 5.4 GB on Mistral Small 3.2 (estimate, section 5.1).
- Holo3-35B-A3B, an Apache-2.0 computer-use fine-tune of Qwen3.5-35B-A3B with 3B active parameters, scores 82.6 in the official OSWorld-Verified results (78.2 in a second run there, 77.8 in its own card) against 86.0 for the proprietary general-purpose model ranked first [benchmarks-34] [benchmarks-55]. It shows that a 3B-active model can operate a user interface; SparkTest reads DOM snapshots as text, not screenshots, so this is indirect evidence.
- Italian-first and European models exist (Almawave Velvet2-3B, EuroLLM-22B, Apertus 1.5) [models-137] [models-132][models-133]. Velvet2-3B was not shortlisted because its configuration is not public (gated repository), so its memory needs cannot be derived [models-137].

For your IT team: on a single device the practical choice is one of the 3B-active MoE models of section 6.1; larger models need a 96 GB card or a multi-GPU server; check table 3.3 before choosing.

4. The hardware

Declared specifications from the manufacturers; measured values labelled as such. Prices are indicative (observed): EU street prices from German or Austrian price comparison on 16 November 2026, VAT included; US prices are list prices without sales tax unless stated; no currency conversion. Card prices exclude the host computer. Detailed tables: companion note `fonti/hardware.md` and data `dati/hardware.json`.

4.1 Class 1: compact unified-memory systems

PLATFORM	MEMORY	BANDWIDTH, DECLARED (MEASURED)	AI COMPUTE, DECLARED (MEASURED)	POWER	INDICATIVE PRICE	SOURCES
NVIDIA GB10, 128 GB (DGX Spark and partner systems)	128 GB LPDDR5X, shared by CPU and GPU	273 GB/s (161 to 230 GB/s idle, about 40 to 105 GB/s while inference runs; community)	FP4 1,000 TFLOPS with sparsity (about 486 to 511 TFLOPS dense NVFP4; community)	140 W chip, 240 W supply	from EUR 5,999 (ASUS Ascent GX10, 1 TB) to EUR 11,383.97 (DGX Spark Founders Edition, 4 TB), street DE; DGX Spark US list USD 6,950 from 2026-10-02	[hardware-1][hardware-2][hardware-15][hardware-16][hardware-17][hardware-9][hardware-8][hardware-4][hardware-5]
NVIDIA GB10, 64 GB (partner systems only, from 2026-10-23)	64 GB LPDDR5X	273 GB/s (press)	as above	as above	from USD 4,999 (expected retail); ASUS Ascent GX10 64 GB from EUR 6,199 (2 offers), street DE, above the 128 GB system on the same day	[hardware-4][hardware-5][hardware-10]
AMD Ryzen AI Max+ 395	up to 128 GB LPDDR5X-8000	256 GB/s (derived from bus and speed; launch figure reported by the press)	GPU FP16 up to 60 TFLOPS; NPU up to 50 TOPS	55 W default, 45 to 120 W	from EUR 3,899.96 (GMKtec EVO-X2, 128 GB), street DE; Framework Desktop 128 GB USD 1,999 at launch, USD 3,449 on 2026-09-30	[hardware-18][hardware-22][hardware-21]
AMD Ryzen AI Max+ PRO 495	192 GB LPDDR5X-8533	273 GB/s	GPU FP16 up to 61 TFLOPS	120 W platform	USD 6,799 to 7,449 (pre-order list, US, 2026-09-30; first systems ship November 2026)	[hardware-19][hardware-21]
NVIDIA Jetson AGX Thor	128 GB LPDDR5X	273 GB/s	FP4 2,070 TFLOPS with sparsity	40 to 130 W	developer kit USD 3,499, raised to USD 5,499 in July 2026	[hardware-25][hardware-26]
NVIDIA N1X ("RTX Spark")	up to 128 GB LPDDR5X-9400	about 300 GB/s (press)	not stated on a comparable basis	100 W (Dev Box)	Surface RTX Spark Dev Box USD 5,999.99, US only	[hardware-27][hardware-28]

Usage terms: the DGX Spark Founders Edition warranty lasts one year, is void for "Enterprise Use" (which the warranty page ties to use in data centres or clusters) and cannot be transferred [hardware-6]; the warranty terms of partner GB10 systems and of Ryzen AI Max+ systems for use at a customer site were not researched. Jetson AGX Thor is a robotics module and N1X a Windows on Arm PC platform; both are listed for completeness and not compared further.

4.2 Class 2: workstation GPUs

CARD	MEMORY	BANDWIDTH	FP8 DENSE / SPARSE (TFLOPS), DECLARED	BOARD POWER	STREET PRICE DE/AT, 2026-11-16	SOURCES
NVIDIA RTX PRO 6000 Blackwell Workstation Edition	96 GB GDDR7 ECC	1,792 GB/s	1,007.6 / 2,015.2	600 W	from EUR 16,290	[hardware-32][hardware-34][hardware-47]
NVIDIA RTX PRO 6000 Blackwell Max-Q	96 GB GDDR7 ECC	1,792 GB/s	877.9 / 1,755.7	300 W	from EUR 16,489.98	[hardware-33][hardware-34][hardware-48]
NVIDIA RTX PRO 5000 Blackwell, 72 GB	72 GB GDDR7 ECC	1,344 GB/s	not published (FP4 with sparsity 2,064)	300 W	from EUR 11,999	[hardware-36][hardware-37][hardware-50] [hardware-92]
NVIDIA RTX PRO 5000 Blackwell, 48 GB	48 GB GDDR7 ECC	1,344 GB/s	not published (FP4 with sparsity 2,064)	300 W	from EUR 8,119	[hardware-36][hardware-37][hardware-49]
NVIDIA RTX PRO 4500 Blackwell	32 GB GDDR7 ECC	896 GB/s	not published	200 W	from EUR 4,559	[hardware-38][hardware-51]
NVIDIA RTX PRO 4000 Blackwell	24 GB GDDR7 ECC	672 GB/s	not published (FP4 with sparsity 1,290)	145 W	from EUR 2,780.97	[hardware-39][hardware-52]
NVIDIA GeForce RTX 5090	32 GB GDDR7	1,792 GB/s	838 / 1,676 (FP16 accumulate)	575 W	from EUR 7,399.99 (MSI variants page; EU list EUR 2,229, March 2025)	[hardware-42][hardware-44][hardware-45]
AMD Radeon AI PRO R9700	32 GB GDDR6	640 GB/s	383 / 766	300 W	from EUR 2,269	[hardware-54][hardware-57]
AMD Radeon PRO W7900	48 GB GDDR6 ECC	864 GB/s	not published (INT8 123 TOPS)	295 W	no offers in DE; USD 3,999 at launch (2023)	[hardware-55][hardware-58]
Intel Arc Pro B70	32 GB GDDR6	608 GB/s	not published (INT8 367 TOPS)	160 to 290 W	from EUR 1,659	[hardware-59][hardware-60][hardware-61]
Intel Arc Pro B60	24 GB GDDR6	456 GB/s	not published (INT8 197 TOPS)	120 to 200 W	from EUR 889	[hardware-59][hardware-62]

Usage terms: the GeForce driver licence states that GeForce software "is not licensed for datacenter deployment" [hardware-43]; AMD states that Radeon PRO and Radeon AI PRO cards are designed for workstations, including rack-mounted ones, and "not designed nor validated for datacenter usage" [hardware-54] [hardware-55]. A deskside NVIDIA DGX Station (252 GB HBM3e plus 496 GB LPDDR5X) has no list price; a reviewer heard figures of USD 100,000 to 125,000 [hardware-30][hardware-31].

4.3 Class 3: data-centre accelerators and servers

ACCELERATOR OR NODE	MEMORY PER GPU	BANDWIDTH PER GPU	FP8 DENSE / SPARSE (TFLOPS), DECLARED	POWER	INDICATIVE PRICE	SOURCES
NVIDIA L4 (PCIe)	24 GB GDDR6	300 GB/s	242.5 (derived) / 485	72 W	from EUR 3,619, street DE	[hardware-63][hardware-82]
NVIDIA L40S (PCIe)	48 GB GDDR6	864 GB/s	733 / 1,466	350 W	from EUR 9,555, street	[hardware-64][hardware-83]
NVIDIA RTX PRO 6000 Server Edition (PCIe)	96 GB GDDR7	1,597 GB/s	not published (FP4 "4 PFLOPS", that is 4,000 TFLOPS, sparsity basis not stated)	up to 600 W	from EUR 20,772.07 per card, street AT	[hardware-35][hardware-84]
NVIDIA H200 NVL (PCIe)	141 GB HBM3e	4.8 TB/s	1,670.5 (derived) / 3,341	up to 600 W	from EUR 38,900 per card, street	[hardware-66][hardware-86]
8 x NVIDIA H100 SXM (HGX)	80 GB HBM3	3.35 TB/s	1,979 (derived) / 3,958	up to 700 W per GPU	USD 195,000 fair market value, used node (US, October 2026)	[hardware-65][hardware-87]

ACCELERATOR OR NODE	MEMORY PER GPU	BANDWIDTH PER GPU	FP8 DENSE / SPARSE (TFLOPS), DECLARED	POWER	INDICATIVE PRICE	SOURCES
8 x NVIDIA H200 SXM (HGX)	141 GB HBM3e	4.8 TB/s	1,979 (derived) / 3,958	up to 700 W per GPU	about USD 340,000 new, volume order (US, October 2026)	[hardware-66][hardware-87]
8 x NVIDIA B200 (HGX)	180 GB HBM3e (derived)	8 TB/s (derived)	4,500 / 9,000	about 14.3 kW per DGX B200 system	USD 360,000 to 470,000 new (US, October 2026)	[hardware-67][hardware-68] [hardware-87][hardware-88]
8 x NVIDIA B300 (HGX)	2.1 TB per 8 GPUs (about 262 GB each, derived)	not listed	4,500 / 9,000	about 14 kW per DGX B300 system	USD 400,000 to 520,000 (US, October 2026)	[hardware-68][hardware-69] [hardware-88][hardware-89]
8 x AMD Instinct MI300X	192 GB HBM3	5.3 TB/s	2,614.9 / 5,229.8	750 W peak	not found	[hardware-72]
8 x AMD Instinct MI355X	288 GB HBM3E	8 TB/s	5,000 / 10,100	1,400 W	not found	[hardware-75]
Intel Gaudi 3	128 GB HBM2e	3.7 TB/s	1,678 (white paper) or 1,835 (launch coverage)	up to 900 W (OAM)	USD 125,000 for 8 with baseboard (2024 list, "for modeling only"); 2026 price not found	[hardware-77][hardware-91] [hardware-79]

The 2026 parts announced or shipping to cloud providers (NVIDIA Rubin with 288 GB HBM4 and 22 TB/s, preliminary; AMD Instinct MI455X with 432 GB HBM4 and 23.3 TB/s) and rack-scale systems (GB300 NVL72 at USD 8.0 to 8.6 million installed) are outside the scope of a SparkTest deployment [hardware-68] [hardware-71][hardware-76][hardware-89].

4.4 What the hardware tables imply

- Class 1 gives 64 to 192 GB in one box at 256 to 273 GB/s; the cards of table 4.2 give 24 to 96 GB per card at 456 to 1,792 GB/s, that is 1.7 to 6.6 times the 273 GB/s of GB10; in class 3 the 8-GPU nodes start at 80 GB and 3.35 TB/s per GPU, and the 4 x RTX PRO 6000 Server Edition configuration has 96 GB and 1,597 GB/s per GPU [hardware-1][hardware-18][hardware-32][hardware-62][hardware-65][hardware-35].
- Bandwidth bounds the generation speed of one request, memory bounds model size and concurrent context, and low-precision compute bounds prompt reading and batches. Declared compute is comparable across vendors only on the same dense or sparse basis and number format; NVIDIA states that its sparse figures are twice the dense ones for FP8, FP16 and INT8 [hardware-63][hardware-67][hardware-68].
- On a unified-memory system the browser automation runs on the same memory as the model; measured GB10 bandwidth falls to about 40 to 105 GB/s on a tool running beside inference [hardware-16][hardware-17]. The effect on SparkTest's call times was not measured.

5. Fit and speed

This section compares the calculator's estimates with public measurements. Estimates use the formulas of section 9; public measurements are labelled by who measured them. Source data for every estimate: `out/tabelle.md` and `out/tabelle.json`.

5.1 Memory per model (estimate)

Weights = parameters held in memory x bits per weight / 8, in GB (10^9 bytes). Precisions: BF16 = 16 bits, 8-bit = 8.5 bits (llama.cpp Q8_0 rate [method-1]), 4-bit = 4.5 bits (NVFP4 rate [method-2]; GGUF Q4_K_M files are 4.89 bits [method-1], about 9% larger). KV cache in BF16 for one 32k session, from the configuration of each model; an FP8 cache halves it.

MODEL	PARAMETERS HELD IN MEMORY (B)	ACTIVE (B)	BF16 (GB)	8-BIT (GB)	4-BIT (GB)	KV, ONE 32K SESSION (GB)	KV, NINE SESSIONS (GB)
Qwen3.6-35B-A3B	36.0	3.0	71.9	38.2	20.2	0.67	6.0
Qwen3.8-27B	27.8	27.8	55.6	29.5	15.6	2.15	19.3
Qwen3.5-122B-A10B	125.1	10.0	250.2	132.9	70.4	0.81	7.2
Qwen3-Coder-Next	79.7	3.0	159.3	84.6	44.8	0.81	7.2
gpt-oss-20b	20.9	3.6	41.8	22.2	11.8	0.81	7.3
gpt-oss-120b	116.8	5.1	233.7	124.1	65.7	1.21	10.9
Gemma 4 26B-A4B	25.8	3.8	51.6	27.4	14.5	0.88	7.9
Gemma 4 31B	31.3	31.3	62.5	33.2	17.6	3.52	31.7
Muse Glimmer 30B	29.8	29.8	59.6	31.6	16.8	0.52	4.7
Nemotron 3.5 Lightning 30B-A3B	31.6	3.0	63.2	33.6	17.8	0.20	1.8
Nemotron 3.5 Bolt 74B-A9B	74.9	9.0	149.8	79.6	42.1	0.27	2.4
Nemotron 3.5 Super VL 120B-A12B	124.5	12.0	248.9	132.2	70.0	0.27	2.4
Mistral Small 4 119B	119.4	6.5	238.8	126.9	67.2	0.75	6.8
Mistral Small 3.2 24B	24.0	24.0	48.0	25.5	13.5	5.37	48.3
Ministral 3 14B	14.0	14.0	27.9	14.8	7.8	5.37	48.3
Granite 4.2 30B	29.3	29.3	58.6	31.1	16.5	8.59	77.3
Laguna XS 2.1	33.4	3.0	66.9	35.5	18.8	1.41	12.6
GLM-4.7-Flash	31.2	3.0	62.4	33.2	17.6	1.77	16.0
EuroLLM-22B	22.6	22.6	45.3	24.1	12.7	7.25	65.2
Qwen3.8-Flash-Next	180.0	6.0	360.0	191.3	101.3	0.81	7.2
GLM-5.3-Flash	321.3	18.0	642.6	341.4	180.7	not derived	not derived

MODEL	PARAMETERS HELD IN MEMORY (B)	ACTIVE (B)	BF16 (GB)	8-BIT (GB)	4-BIT (GB)	KV, ONE 32K SESSION (GB)	KV, NINE SESSIONS (GB)
DeepSeek-V4-Flash-0731	304.2	13.0	608.4	323.2	171.1	not derived	not derived
Qwen3.5-397B-A17B	403.4	17.0	806.8	428.6	226.9	1.01	9.1
Nemotron 3 Ultra 550B-A55B	560.5	55.0	1,121.0	595.6	315.3	0.40	3.6
Mistral Large 3 675B	675.0	41.0	1,350.0	717.2	379.7	2.30	20.7
DeepSeek-V4.1-Flash-1017	763.2	16.0	1,526.4	810.9	429.3	0.03 (declared 890 bytes per token)	0.3
GLM-5.3	753.3	40.0	1,506.7	800.4	423.7	2.94	26.5
Kimi K3.1 Pro	2,780.1	104.0	5,560.2	2,953.9	1,563.8	0.91	8.2

Parameters held in memory are the Hugging Face safetensors counts (they include vision encoders and multi-token-prediction layers) [models-12][models-15] [models-39]; active parameters are declared, except GLM-5.3, whose card gives no count (40B active is declared for GLM-5) [models-85][models-82]. Linear-attention and state-space layers keep a fixed state per session that is not included. The calculator's KV figures match the derivation in the companion model note for every model where it exists.

5.2 Fit matrix (estimate)

Cell: the first of BF16, 8-bit and 4-bit at which the weights plus **nine** 32k sessions (size L) fit beside the reserve (16 GB on a unified-memory system, 10% of card memory on a GPU; section 9). "4-bit (n)": no precision holds nine sessions; n concurrent 32k sessions fit at 4 bits. "no": not even one session at 4 bits. All 32 GB cards behave alike in this test, as do all 48 GB and all 24 GB cards. A size needs 3 (S), 5 (M) or 9 (L) concurrent sessions: a cell such as "4-bit (6)" serves sizes S and M only.

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

MODEL	GB10 64 GB	GB10 128 GB	RYZEN AI MAX+ 395, 128 GB	RYZEN AI MAX+ PRO 495, 192 GB	RTX PRO 6000, 96 GB	RTX PRO 5000, 72 GB	48 GB CARD	32 GB CARD	24 GB CARD
Mistral Small 4 119B	no	4-bit	4-bit	8-bit	4-bit	no	no	no	no
Mistral Small 3.2 24B	4-bit (6)	BF16	BF16	BF16	8-bit	4-bit	4-bit (5)	4-bit (2)	4-bit (1)
Ministral 3 14B	4-bit (7)	BF16	BF16	BF16	BF16	8-bit	4-bit (6)	4-bit (3)	4-bit (2)
Granite 4.2 30B	4-bit (3)	8-bit	8-bit	BF16	4-bit (8)	4-bit (5)	4-bit (3)	4-bit (1)	no
Laguna XS 2.1	4-bit	BF16	BF16	BF16	BF16	8-bit	4-bit	4-bit (7)	4-bit (1)
GLM-4.7-Flash	4-bit	BF16	BF16	BF16	BF16	8-bit	4-bit	4-bit (6)	4-bit (2)
EuroLLM-22B	4-bit (4)	BF16	BF16	BF16	4-bit	4-bit (7)	4-bit (4)	4-bit (2)	4-bit (1)

Data-centre configurations (`*` = KV cache not derived, weights only):

MODEL	L40S, 48 GB	RTX PRO 6000 SERVER, 96 GB	H200 NVL, 141 GB	4 X RTX PRO 6000 SERVER, 384 GB	8 X H100, 640 GB	8 X H200, 1,128 GB	8 X B200, 1,440 GB	8 X MI355X, 2,304 GB
Qwen3.8-Flash-Next	no	no	4-bit	8-bit	BF16	BF16	BF16	BF16
GLM-5.3-Flash	no	no	no	8-bit*	8-bit*	BF16*	BF16*	BF16*
DeepSeek-V4-Flash-0731	no	no	no	8-bit*	8-bit*	BF16*	BF16*	BF16*
Qwen3.5-397B-A17B	no	no	no	4-bit	8-bit	BF16	BF16	BF16
Nemotron 3 Ultra 550B-A55B	no	no	no	4-bit	4-bit	8-bit	BF16	BF16
Mistral Large 3 675B	no	no	no	no	4-bit	8-bit	8-bit	BF16
DeepSeek-V4.1-Flash-1017	no	no	no	no	4-bit	8-bit	8-bit	BF16
GLM-5.3	no	no	no	no	4-bit	8-bit	8-bit	BF16
Kimi K3.1 Pro	no	no	no	no	no	no	no	4-bit
gpt-oss-120b	no	4-bit	4-bit	BF16	BF16	BF16	BF16	BF16
Qwen3.6-35B-A3B	4-bit	BF16	BF16	BF16	BF16	BF16	BF16	BF16

Nine sessions of 32k tokens is a conservative test: SparkTest's typical prompts are 4,000 to 11,000 tokens (section 1.2), and engines with a paged KV cache allocate what each session uses.

5.3 Speed: estimates beside public measurements

Single-request generation (decode) is estimated as bandwidth / bytes read per token x an efficiency factor calibrated on public measurements (section 9.3); prompt reading (prefill) as the effective compute measured on the same hardware / (2 x active parameters). The rows below also set the calibration (they are in-sample), except the Ollama value (42.1) and the vLLM INT4 value (32.8), which stay out of it; so the table shows the spread around the medians, not an independent validation. Measured values are in tok/s, single request, empty or near-empty context unless stated. Estimates here use the bits of the measured file and the depth of the measurement; section 6 uses 4.5 or 8.5 bits per weight at the depth of a codify call, so the two tables can differ for the same pair (for example gpt-oss-120b on GB10: 55.1 here, 50 in section 6).

Decode:

MODEL	HARDWARE	RUNTIME, QUANTIZATION	MEASURED (WHO)	ESTIMATE	SOURCES
gpt-oss-120b	DGX Spark (GB10)	llama.cpp build 7941, MXFP4	58.7 (runtime developer)	55.1	[measurements-2]
gpt-oss-120b	DGX Spark (GB10)	vLLM 0.26.1rc1 / Ollama	60.7 / 42.1 (community)	55.1	[measurements-12]
gpt-oss-20b	DGX Spark (GB10)	llama.cpp build 7941, MXFP4	83.4 (runtime developer)	73.1	[measurements-2]
gpt-oss-120b	Ryzen AI Max+ 395	llama.cpp, ROCm / Vulkan	51.8 / 56.6 (community)	51.6	[measurements-20]
Qwen3.6-35B-A3B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL, ROCm / Vulkan	51.5 / 60.4; 51.3 / 60.7 (community)	73.5 to 75.0	[measurements-20] [measurements-21]
Gemma 4 26B-A4B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	48.5 / 54.7 (community)	56.9	[measurements-20]
Qwen3.8-27B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	11.9 (community)	11.9	[measurements-21]
Gemma 4 31B	Ryzen AI Max+ 395	llama.cpp, UD-Q4_K_XL	10.6 / 11.5 (community)	11.4	[measurements-20]
Qwen3.5-122B-A10B	Ryzen AI Max+ 395	llama.cpp, UD-Q5_K_XL	20.1 / 22.3 (community)	19.4	[measurements-20]
gpt-oss-120b	RTX PRO 6000 Workstation	llama.cpp, MXFP4	196 (runtime developer)	179	[measurements-5]
gpt-oss-20b	GeForce RTX 5090	llama.cpp, MXFP4	283 (runtime developer)	238	[measurements-5]
Qwen3.6-35B-A3B	Radeon AI PRO R9700	llama.cpp server, Vulkan / ROCm	114.2 / 76.4 (community)	91.2	[measurements-21]
Qwen3.8-27B	Radeon AI PRO R9700	llama.cpp server, Vulkan / ROCm	29.3 / 26.6 (community)	23.9	[measurements-21]
Qwen3.8-27B	Arc Pro B70	llama.cpp Vulkan / SYCL; vLLM INT4	22.7 / 12.4; 32.8 (community)	22.7	[measurements-21]

Prefill:

MODEL	HARDWARE	MEASURED (WHO)	ESTIMATE	SOURCES
gpt-oss-120b	DGX Spark (GB10)	2,444 (llama.cpp build 7941); 1,919 (build 6989); 1,725 (NVIDIA); 1,169 (Ollama)	1,933	[measurements-2][measurements-3] [measurements-9][measurements-11]
gpt-oss-20b	DGX Spark (GB10)	4,506; 3,714; 3,670; 3,224 (same sources)	2,738	[measurements-2][measurements-3] [measurements-9][measurements-11]
Qwen3.6-35B-A3B	Ryzen AI Max+ 395	1,120 / 1,114 (pp512); 1,382 / 1,219 (server, 2,048-token chunk)	1,120	[measurements-20][measurements-21]
Qwen3.8-27B	Ryzen AI Max+ 395	409	363	[measurements-21]
gpt-oss-120b	RTX PRO 6000 Workstation	5,518	6,826	[measurements-5]
Qwen3.6-35B-A3B	Radeon AI PRO R9700	3,793 / 3,967	3,793	[measurements-21]
Qwen3.8-27B	Radeon AI PRO R9700	1,022 / 1,080	1,051	[measurements-21]
Qwen3.8-27B	Arc Pro B70	1,742 (vLLM); 1,255 (SYCL); 706 (Vulkan)	1,255	[measurements-21]

The archive of [measurements-21] holds other runs of the same pairs with different values (for example vLLM INT4 on the Arc Pro B70: decode 31 to 41 tok/s, prefill 1,740 to 2,211 tok/s); the tables show one run per engine.

What the comparison shows:

- Generation tracks bandwidth: on the compact systems the measured values sit at 38% to 91% of the bandwidth limit, on the cards at 21% to 82%; small-active MoE models on cards sit at the bottom of that range (section 9.3).
- The estimate is high by 21% to 46% for Qwen3.6-35B-A3B on Ryzen AI Max+ 395. For dense models it falls within about 20% of most measurements; the exceptions are engine effects (on the Arc Pro B70 the same model generates 12.4 to 32.8 tok/s depending on the engine).
- Software moves the result as much as hardware: llama.cpp prefill for gpt-oss-120b on DGX Spark rose 27% between two builds [measurements-2][measurements-3]; Ollama generated about 30% slower than llama.cpp for gpt-oss-120b on the same box [measurements-11][measurements-2]; on the Arc Pro B70 three engines span 706 to 1,742 tok/s of prefill [measurements-21]; on Ryzen AI Max+ 395 five software backends span 234 to 653 tok/s of gpt-oss-120b prefill at 32k depth [measurements-20].
- Context depth costs 22% to 30% of decode speed at 32k on the compact systems for gpt-oss-120b, across two llama.cpp builds on DGX Spark and five backends on Ryzen AI Max+ 395 [measurements-2] [measurements-3][measurements-20].

5.4 Concurrency

Several requests at once share the weights read from memory, so the total generation rate rises with the number of requests but each request slows down. Measured: gpt-oss-120b on DGX Spark with llama.cpp went from 55 tok/s with one request to 139 tok/s in total with 8 (17.4 per request) [measurements-2]; with SGLang from 51 to 117 [measurements-7]. In a same-harness test vLLM took gpt-oss-120b from 60.7 tok/s to 116 to 153 tok/s with 8 concurrent requests, while Ollama gave 42.5 tok/s with 8 requests against 42.1 alone [measurements-12]. On an RTX PRO 6000, gpt-oss-120b with SGLang went from 208 to 640 tok/s at 8 requests [measurements-7]. Dense models scale almost linearly up to 8 requests (Gemma 3 27B FP8 with SGLang: 3.8 to 27.8 tok/s on DGX Spark, 22.6 to 158.6 on RTX PRO 6000) [measurements-7].

The calculator models the total generation rate at c requests as the single-request rate x c^beta, with beta fitted on these series: 0.44 for MoE (0.35 to 0.66 over 9 series) and 0.96 for dense models (0.94 to 1.10 over 4 series). In the same-harness test vLLM also took Qwen3-30B-A3B from 77.1 to 313.4 tok/s at 8 requests (beta 0.67), against beta 0.31 to 0.44 for gpt-oss-120b [measurements-12]: continuous batching can scale some MoE models better than the median used.

5.5 Data-centre throughput (measured, no estimate made)

HARDWARE	MODEL	MEASURE	VALUE	WHO MEASURED	SOURCES
H100 / H200 / B200 (per GPU)	gpt-oss-120b, MXFP4, vLLM	total tok/s per GPU at 50 tok/s per user	8,973 / 9,319 / 44,489	independent analyst (InferenceX)	[measurements-32] [measurements-33] [measurements-34]
MI300X / MI355X (per GPU)	gpt-oss-120b, MXFP4, vLLM	same	8,067 / 38,497	independent analyst	[measurements-35] [measurements-36]
H100 / H200 / B200 / MI355X (per GPU)	Qwen3.5-397B-A17B, SGLang	same	1,689 / 1,810 / 9,774 / 8,132	independent analyst	[measurements-40] [measurements-41] [measurements-42] [measurements-43]
8 x RTX PRO 6000 Server Edition	gpt-oss-120b	MLPerf v6.1 offline output tok/s per GPU (other submissions for the same platform are 2% to 5% lower)	about 1,967	MLPerf submitter	[measurements-26] [measurements-28]
8 x B200 / 8 x MI355X	gpt-oss-120b	same	11,436 / 15,227	MLPerf submitters	[measurements-26]
4 x RTX PRO 6000 Server Edition	DeepSeek-V4-Flash, SGLang	output tok/s, 1 / 16 requests	39.1 / 349.1	service provider	[measurements-51]
2 x DGX Spark	Qwen3.8-Flash-Next, NVFP4	single-request decode; time to first token, 32k prompt	29.6 tok/s; 15.9 s	community	[measurements-13]

SparkTest's need at size L is nine requests at about 25 to 32 tok/s each (section 1.4), which is a small fraction of these figures. On a multi-GPU server the question is cost, power and licence, not speed.

5.6 End-to-end agent measurements

Two public tests replay agent work end to end; speculative decoding is on in all AgentPerf-Local entries and in the MLPerf entries on DGX Spark and Ryzen AI Max+ 395:

TEST	HARDWARE	MODEL	RESULT	WHO MEASURED	SOURCES
MLPerf Inference v6.1 Edge Agentic (1,007 turns, 32k context per turn, thinking off)	DGX Spark / Ryzen AI Max+ 395 desktop	Qwen3.6-27B, NVFP4 with MTP	3.81 s / 7.06 s mean per turn; 29.5 / 17.9 tok/s (from mean time per output token)	MLPerf submitter (the maker of the inference engine used)	[measurements-26]
same	Jetson AGX Thor (NVIDIA submission) / RTX PRO 4500 Server Edition (32 GB)	Qwen3.6-27B, FP4 / INT4	1.47 s / 2.76 s mean per turn	MLPerf submitters	[measurements-26]
Artificial Analysis AgentPerf-Local (8 tasks, 168 turns, reasoning on)	DGX Spark / Ryzen AI Halo (Ryzen AI Max+ 395, 128 GB) / RTX 5090	Qwen3.6-35B-A3B, 4-bit	7.3 / 11.7 / 2.0 minutes	independent analyst	[measurements-25] [measurements-24]
same	same	Qwen3.8-27B, 4-bit	17.3 / 34.5 / 4.9 minutes	independent analyst	[measurements-25] [measurements-24]

Both tests rank the systems the same way as the calculator. DGX Spark finished AgentPerf-Local 1.6 to 2.0 times sooner than the Ryzen AI Halo on these two models; the article reports 1.4 to 1.7 times on three of its four models [measurements-25][measurements-24]. The AgentPerf-Local runs used reasoning (the results page labels them "Qwen3.6 35B A3B (Reasoning)" and "Qwen3.8 27B (Xhigh)"), while SparkTest switches it off; the MLPerf workload runs with thinking off [measurements-29]. A press summary of MLPerf v6.1 gives 19.63 tok/s for the Ryzen AI Max+ 395 entry, which matches no field of the official file (output tokens per second over the run: 10.22) [measurements-29][measurements-26]; this study uses the official files. Both tests also show that speculative decoding (MTP) lifts a dense 27B model on a compact system to about 29.5 tok/s per request [measurements-26], well above the 6 to 12 tok/s measured without it for dense 24B to 32B models on compact systems [measurements-11][measurements-20][measurements-21]; the calculator does not include speculative decoding, whose gain shrinks under concurrency [measurements-13].

5.7 Measured on SparkTest's own tasks

The public tests above replay other agents' work. On 17 November 2026 SparkTech ran SparkTest's model qualification kit (section 7.5) on two models of the class recommended in section 6.1: Qwen3.6-35B-A3B (Q4_K_M, 4.74 bits per weight) and gpt-oss-20b (MXFP4, 4.63 bits per weight), served by llama.cpp v1.0.0 (build b12359) on a test server that is not a SparkNode: a 2016 processor (Intel Xeon E5-2630 v4, 10 cores, AVX2), 256 GB of DDR4 memory, no GPU, shared with other services, the model on 10 threads at minimum priority. The kit sends each model the agent's own prompts for eight tasks (three codify rounds, a whole plan, two repairs, a diagnosis and an API analysis, in English and Italian) and judges every answer with the agent's own checks. Full document, with every answer kept: [method-6]. Values in this section are measured by SparkTech unless marked as estimates.

MEASURE	QWEN3.6-35B-A3B Q4_K_M	GPT-OSS-20B MXFP4	WHAT THE KIT ASKS
Required checks (answer, format, the agent's validation)	22/23	23/23	all of them
Quality checks (the expected behaviour)	24/25	25/28	the behaviour each task asks for
Answers constrained by a schema that pass the kit's schema check	7/7	7/7	all of them
Legitimate repair (a renamed button) accepted by the scenario check	yes	yes	yes
Repair shown a page with a value other than the expected one	value changed, repair rejected	value changed, repair rejected	value kept
Diagnosis in Italian, three sections, evidence cited	yes, 118 words	yes, 119 words	60 to 100 words, 120 at most

MEASURE	QWEN3.6-35B-A3B Q4_K_M	GPT-OSS-20B MXFP4	WHAT THE KIT ASKS
Defects planted in the last release found by the API analysis	2 of 2	1 of 2	2 of 2
Calls within the agent's time limits on this server	1 of 8	1 of 8	8 of 8
Whole call, median, on this server	100.7 s	108.5 s	

What it shows for this study:

- **The answers carry over to a node, the times do not.** Answers depend on the model, its quantization, the engine and the prompt, not on the speed of the machine. On this processor prompt reading ran at about 35 tok/s and took 56% to 90% of each call; public measurements of the same models on class 1 hardware read prompts 31 to 124 times faster (section 5.3). Generation reached 31% (Qwen3.6) and 52% (gpt-oss-20b) of the ceiling set by the declared 68.3 GB/s, one machine under load with a declared, not measured, bandwidth. The speed of a node is measured on its own hardware (section 7.5).
- **The agent's checks do the decisive work.** The engine's grammar made every constrained answer follow its schema; the agent's validator rejected the one invalid plan (Qwen3.6 put the expected text of two assertions in the wrong field); gpt-oss-20b filled a negative sign-in test with the credential placeholders instead of the test's own user, which would end in a visible failure. Shown a page whose value differs from the one the test expects, both models changed the expected value and the scenario check rejected both repairs, so the regression stayed visible. The repair prompt measured showed the model the test's title but not its description, where the value is a requirement, and asked it to change the failing assumption: the result measures that prompt as much as the models. The prompt will give the model the description and an explicit rule never to change an expected value; the scenario check stays the limit that does not depend on the model.
- **Italian works on both models** in the diagnosis and the API analysis. Qwen3.6 found both defects planted in the API statistics, gpt-oss-20b one, and gpt-oss-20b stated a database detail that the statistics do not contain; the kit checks invented endpoints, not every technical claim.
- **Reasoning must be off.** With the server's default, Qwen3.6 reasoned until the probe's 1,024-token cap, for 89 s, without an answer; `chat_template_kwargs` switched reasoning off (answer in 1.2 s). The measured version of the agent's request did not carry that field, so the runs switched reasoning off on the server; the agent sends it since 17 November 2026 (section 7.3).
- **Token assumptions hold, except for API answers.** Codify, repair and diagnosis prompts had 3.4 to 4.3 characters per token (section 1.2 assumes 4.0 for English and 3.5 for Italian); the API statistics prompt had 2.6 to 2.8, 23% to 35% more tokens than assumed. The two API analyses were 881 and 700 answer tokens against the 418 of section 1.2, which raises the API-insights estimate for Qwen3.6 at 4 bits to about 12 s on GB10 and 16 s on Ryzen AI Max+ 395 (estimate), far below its 150 s limit.

Limits: one sample per task; one request at a time, so no concurrency; plans validated, not executed in a browser; 4-bit weights only; reasoning set on the server; the agent's prompts have changed since the measurement (6 of the 8 prompts are 136 characters longer). It is a first qualification of two models, not a benchmark, and it says nothing about the speed of the hardware of section 6.

6. Configurations compared and recommended

Per-call times below are estimates for a burst in which all concurrent calls of a size arrive together, which is an upper estimate; isolated calls are much shorter. Two prompt cases are checked: typical prompts (codify round of 4,941 tokens) and application documentation at the 24,000-character cap (codify round of 11,093 tokens); repair and diagnosis keep their typical prompts in both cases (section 1.4). "Usable" is the largest size whose bursts meet the agent's hard limits in both cases. Decode is given at the depth of a typical codify call (about 5,000 tokens) with 4.5 or 8.5 bits per weight; GGUF 4-bit files (4.9 to 5.2 bits per weight) decode about 8% to 13% slower than this table on llama.cpp (section 5.3). "Release night" is the number of maintained tests whose release-night load (high bound per test) fits in a 6-hour night at the concurrency of each size. Full tables for every pair: `out/tabelle.md`, tables 7, 10, 11 and 12.

A size within the hard limits but not the targets still works, more slowly: codify rounds longer than 15 s use up the 4-minute exploration budget sooner, so the agent falls back to a one-shot plan more often, and runs wait longer for their repairs [workload-serving-4]. A call that passes its hard limit is stopped and retried once with a shorter prompt [workload-serving-1].

CONFIGURATION	MODEL, WEIGHTS	32K SESSIONS THAT FIT	PREFILL / DECODE AT CODIFY DEPTH, TOK/S	CODIFY: ALONE / BURST L (S)	CODIFY WITH DOCUMENTATION AT CAP: ALONE / BURST M / BURST L (S)	REPAIR: ALONE / S / M / L (S)	TARGETS / HARD LIMITS, TYPICAL PROMPTS	TARGETS / HARD LIMITS, DOCUMENTATION AT CAP	USABLE	RELEASE NIGHT AT S / M / L CONCURRENCY (TESTS)
GB10, 128 GB	Qwen3.6-35B-A3B, 4-bit	136	3,286 / 85	2.8 / 18.1	4.8 / 20.4 / 35.2	9.5 / 18.8 / 26.1 / 38.5	M / L	S / L	up to L	7,960 / 9,105 / 10,463
GB10, 128 GB	Qwen3.6-35B-A3B, 8-bit	109	3,286 / 46	4.0 / 21.9	5.9 / 23.1 / 39.1	16.6 / 31.9 / 43.5 / 62.6	none / M	none / M	up to M	5,240 / 6,181 / 7,375
GB10, 128 GB	Nemotron 3.5 Lightning, 4-bit	468	3,286 / 89	2.8 / 17.9	4.7 / 20.1 / 34.8	9.2 / 18.3 / 25.4 / 37.6	M / L	S / L	up to L	8,163 / 9,316 / 10,677
GB10, 128 GB	gpt-oss-120b, 4-bit	38	1,933 / 50	4.8 / 30.8	8.2 / 34.7 / 60.0	16.2 / 32.1 / 44.5 / 65.7	none / M	none / M	up to M	4,668 / 5,341 / 6,140
GB10, 128 GB	Mistral Small 4, 4-bit	59	1,517 / 40	6.1 / 38.9	10.2 / 43.7 / 75.8	20.2 / 40.0 / 55.6 / 82.1	none / M	none / M	up to M	3,738 / 4,269 / 4,896
GB10, 128 GB	Qwen3.8-27B, 4-bit	44	701 / 14	15.0 / 72.2	24.0 / 87.9 / 152	55.4 / 68.5 / 80.3 / 103	none / none	none / none	no	2,049 / 2,590 / 3,159
GB10, 64 GB	Qwen3.6-35B-A3B, 4-bit	41	3,286 / 85	2.8 / 18.1	4.8 / 20.4 / 35.2	9.5 / 18.8 / 26.1 / 38.5	M / L	S / L	up to L	7,960 / 9,105 / 10,463
Ryzen AI Max+ 395, 128 GB	Qwen3.6-35B-A3B, 4-bit	136	1,120 / 80	5.8 / 44.5	11.4 / 53.3 / 94.3	12.2 / 26.3 / 38.3 / 60.0	none / L	none / M	up to M	4,479 / 4,845 / 5,230
Ryzen AI Max+ 395, 128 GB	Nemotron 3.5 Lightning, 4-bit	468	1,120 / 83	5.8 / 44.4	11.3 / 52.9 / 93.9	11.9 / 25.8 / 37.6 / 59.0	none / L	none / M	up to M	4,547 / 4,908 / 5,286
Ryzen AI Max+ 395, 128 GB	gpt-oss-120b, 4-bit	38	659 / 47	9.9 / 75.8	19.5 / 90.6 / 160	20.9 / 44.9 / 65.4 / 102	none / S	none / S	up to S	2,630 / 2,845 / 3,072
Ryzen AI Max+ 395, 128 GB	Qwen3.8-27B, 4-bit	44	363 / 13	22.1 / 132	39.2 / 162 / 285	63.7 / 86.7 / 108 / 151	none / none	none / none	no	1,401 / 1,652 / 1,883
Ryzen AI Max+ PRO 495, 192 GB	Qwen3.5-122B-A10B, 8-bit	53	342 / 14	22.5 / 157	40.6 / 182 / 320	61.6 / 126 / 178 / 269	none / none	none / none	no	1,079 / 1,202 / 1,339
RTX PRO 6000, 96 GB	Qwen3.6-35B-A3B, 8-bit	71	11,604 / 151	1.2 / 6.4	1.7 / 6.7 / 11.3	5.1 / 9.7 / 13.2 / 19.0	L / L	L / L	up to L	17,410 / 20,608 / 24,701
RTX PRO 6000, 96 GB	gpt-oss-120b, 4-bit	17	6,826 / 162	1.4 / 8.9	2.4 / 10.0 / 17.2	4.9 / 9.7 / 13.4 / 19.8	L / L	M / L	up to L	15,669 / 18,007 / 20,807
RTX PRO 6000, 96 GB	Qwen3.8-27B, 8-bit	26	3,215 / 40	4.4 / 16.9	6.3 / 20.3 / 34.2	19.0 / 22.2 / 24.9 / 30.0	M / L	S / L	up to L	7,019 / 9,405 / 12,249
RTX PRO 6000, 96 GB	Mistral Small 4, 4-bit	25	5,356 / 132	1.8 / 11.2	3.0 / 12.6 / 21.7	6.1 / 12.1 / 16.8 / 24.7	L / L	M / L	up to L	12,553 / 14,400 / 16,604
RTX PRO 6000, 96 GB	Qwen3.5-122B-A10B, 4-bit	19	3,481 / 86	2.7 / 17.3	4.5 / 19.2 / 33.3	9.4 / 18.5 / 25.6 / 37.7	M / L	S / L	up to L	8,208 / 9,411 / 10,845
RTX PRO 5000, 72 GB	Qwen3.6-35B-A3B, 8-bit	39	5,802 / 113	1.9 / 11.1	3.0 / 12.1 / 20.8	7.0 / 13.6 / 18.7 / 27.2	L / L	M / L	up to L	11,681 / 13,563 / 15,871
RTX PRO 5000, 72 GB	Qwen3.8-27B, 8-bit	16	1,608 / 30	6.8 / 31.8	10.7 / 38.6 / 66.3	26.1 / 31.9 / 37.0 / 46.9	none / L	none / M	up to M	4,493 / 5,735 / 7,070
RTX PRO 5000, 48 GB	Qwen3.6-35B-A3B, 4-bit	34	5,802 / 208	1.4 / 9.5	2.5 / 11.0 / 19.2	4.1 / 8.2 / 11.6 / 17.4	L / L	M / L	up to L	16,971 / 19,062 / 21,447
RTX PRO 5000, 48 GB	Qwen3.8-27B, 4-bit	12	1,608 / 57	5.1 / 29.9	9.0 / 36.7 / 64.4	15.0 / 20.2 / 25.1 / 34.7	S / L	none / M	up to M	6,081 / 7,210 / 8,258
GeForce RTX 5090, 32 GB	Qwen3.6-35B-A3B, 4-bit	12	9,422 / 277	0.9 / 6.1	1.6 / 7.0 / 12.1	3.0 / 5.9 / 8.3 / 12.2	L / L	L / L	up to L	24,656 / 28,004 / 31,914
Radeon AI PRO R9700, 32 GB	Qwen3.6-35B-A3B, 4-bit	12	3,793 / 99	2.5 / 15.6	4.2 / 17.6 / 30.5	8.2 / 16.2 / 22.5 / 33.2	M / L	S / L	up to L	9,234 / 10,558 / 12,126
Radeon AI PRO R9700, 32 GB	Qwen3.8-27B, 4-bit	6	1,051 / 27	8.9 / 47.0	14.9 / 57.4 / 99.8	30.1 / 38.5 / 46.2 / 61.0	none / M	none / M	up to M	3,453 / 4,246 / 5,037
Arc Pro B70, 32 GB	Qwen3.6-35B-A3B, 4-bit	12	4,529 / 94	2.3 / 13.9	3.7 / 15.4 / 26.4	8.4 / 16.4 / 22.6 / 33.0	M / L	S / L	up to L	9,516 / 11,015 / 12,838
Arc Pro B70, 32 GB	Qwen3.8-27B, 4-bit	6	1,255 / 26	8.4 / 40.3	13.4 / 49.1 / 84.6	31.0 / 38.3 / 44.9 / 57.4	none / M	none / M	up to M	3,667 / 4,635 / 5,655
RTX PRO 4000, 24 GB	Nemotron 3.5 Lightning, 4-bit	19	3,626 / 108	2.4 / 15.8	4.1 / 17.9 / 31.2	7.7 / 15.3 / 21.3 / 31.7	M / L	S / L	up to L	9,572 / 10,863 / 12,368

CONFIGURATION	MODEL, WEIGHTS	32K SESSIONS THAT FIT	PREFILL / DECODE AT CODIFY DEPTH, TOK/S	CODIFY: ALONE / BURST L (S)	CODIFY WITH DOCUMENTATION AT CAP: ALONE / BURST M / BURST L (S)	REPAIR: ALONE / S / M / L (S)	TARGETS / HARD LIMITS, TYPICAL PROMPTS	TARGETS / HARD LIMITS, DOCUMENTATION AT CAP	USABLE	RELEASE NIGHT AT S / M / L CONCURRENCY (TESTS)
Arc Pro B60, 24 GB	Nemotron 3.5 Lightning, 4-bit	19	2,215 / 74	3.8 / 25.3	6.6 / 28.9 / 50.4	11.4 / 23.0 / 32.2 / 48.2	S / L	none / L	up to L	6,210 / 7,006 / 7,922
L40S, 48 GB	Qwen3.6-35B-A3B, 4-bit	34	8,242 / 134	1.5 / 8.3	2.3 / 9.0 / 15.2	5.8 / 11.2 / 15.3 / 22.1	L / L	M / L	up to L	14,578 / 17,082 / 20,215
H200 NVL, 141 GB	gpt-oss-120b, 4-bit	50	11,048 / 435	0.7 / 4.9	1.3 / 5.7 / 10.0	2.0 / 4.0 / 5.7 / 8.6	L / L	L / L	up to L	33,953 / 37,936 / 42,427
H200 NVL, 141 GB	Qwen3.8-Flash-Next, 4-bit (licence: legal opinion needed, table 3.3)	31	9,391 / 380	0.8 / 5.8	1.5 / 6.7 / 11.7	2.3 / 4.6 / 6.6 / 9.9	L / L	L / L	up to L	29,279 / 32,662 / 36,463

Targets: codify round 15 s, repair 30 s, diagnosis 30 s; hard limits 60 s, 60 s, 30 s (section 1.4). Prefill and decode are estimates (section 5.3); the prefill estimates for RTX PRO 5000, RTX PRO 4000, Arc Pro B60, L40S and H200 NVL are scaled from declared compute because no prefill measurement was found for those cards (section 9.3). Quality at each precision is not estimated (section 8.3). Rows marked "no" in the "Usable" column are shown for comparison.

6.1 Class 1: compact unified-memory system

- **Hardware:** NVIDIA GB10 systems (64 or 128 GB, 273 GB/s declared) and AMD Ryzen AI Max+ systems (128 GB at 256 GB/s derived, 192 GB at 273 GB/s declared) (section 4.1).
- **Models (estimate):** MoE models with about 3B active parameters at 4 bits: Qwen3.6-35B-A3B, Nemotron 3.5 Lightning 30B-A3B, Laguna XS 2.1, Gemma 4 26B-A4B, gpt-oss-20b. GLM-4.7-Flash fits too, but its card names English and Chinese only [models-79]. All of them hold nine 32k sessions with room to spare, even on 64 GB.
- **Volumes on GB10 (estimate):** with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1, within the targets up to M with typical prompts (repair burst at M 25.4 to 27.8 s against 30 s) and up to S with documentation at the cap (codify burst at M 20.1 to 20.9 s against 15 s); usable up to L (a burst of nine repairs takes 37.6 to 40.9 s against the 60 s limit). With Gemma 4 26B-A4B or gpt-oss-20b, within the targets up to S with typical prompts only; usable up to L (table 12 of `out/tabelle.md`).
- **Volumes on Ryzen AI Max+ 395 (estimate):** isolated calls stay within the targets (codify 5.8 s, repair 12.2 s), but no size does in a burst, because prompt reading is about three times slower than on GB10 (the codify burst at S takes 15.9 s against 15 s). Usable up to M with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1, up to S with Gemma 4 26B-A4B or gpt-oss-20b. Size L is not recommended: with typical prompts the repair burst at L takes 59.0 to 60.0 s against the 60 s limit (62.5 to 73.8 s for Qwen3.6-35B-A3B at the lower decode efficiencies of table 12), and with documentation at the cap the codify burst at L takes about 94 s. The PRO 495 gives the same usable sizes (table 7 of `out/tabelle.md`).
- **Public end-to-end tests** agree on the ranking: DGX Spark finished AgentPerf-Local 1.6 to 2.0 times sooner than the Ryzen AI Halo (Ryzen AI Max+ 395, 128 GB) [measurements-25], and an MLPerf Edge Agentic turn took 3.81 s against 7.06 s [measurements-26] (section 5.6).
- **Margin:** size M within the targets on GB10 rests on the calibrated decode efficiency (0.56), which gpt-oss measurements dominate (11 of the 14 GB10 points). With the efficiency measured for Qwen3.6-35B-A3B on Ryzen AI Max+ 395 (0.38 to 0.46), Qwen3.6-35B-A3B and Laguna XS 2.1 drop to S on GB10, Nemotron 3.5 Lightning at 0.38; the usable size of these three models does not change (table 12). Size M within the targets must be confirmed by the qualification run of section 7.5.
- **Precision:** these sizes need 4-bit weights. At 8 bits GB10 meets no size within the targets and is usable up to M (table 6). The 4-bit model must pass the qualification of section 7.5 on Italian and long prompts (section 8.3).
- **Night capacity is not the limit:** a release night at size-L concurrency fits 7,900 to 10,700 maintained tests on GB10 and 4,000 to 5,300 on Ryzen AI Max+ 395 with the five models above (estimate).
- **Does not do:** size L within the targets; dense 24B to 32B models at the targets (6 to 12 tok/s of generation measured, 13 to 14 estimated; a repair alone takes about 55 to 64 s, estimate), unless speculative decoding is qualified on the chosen engine [measurements-26]; 117B to 125B MoE within the targets: gpt-oss-120b and Mistral Small 4 are usable up to M on GB10 (repair burst at M 44.5 and 55.6 s), and on Ryzen AI Max+ 395 only gpt-oss-120b, up to S (its repair burst at M takes 65.4 s, above the 60 s limit); the 192 GB system opens 8-bit weights for 120B-class models but not the speed to use them at SparkTest's latencies; data-centre models. The DGX Spark Founders Edition warranty excludes "Enterprise Use" [hardware-6].

6.2 Class 2: single workstation GPU

- **Hardware:** one card in a workstation or a rack-mounted workstation: RTX PRO 6000 (96 GB), RTX PRO 5000 (48 or 72 GB), 32 GB cards (RTX PRO 4500, Radeon AI PRO R9700, Arc Pro B70; GeForce RTX 5090 with the licence restriction below) (section 4.2). 24 GB cards (RTX PRO 4000, Arc Pro B60) are a narrower option, described at the end of the volumes.
- **Models (estimate):** with 96 GB, every node-scale candidate in table 3.1: 3B-active MoE at 8 bits or BF16, dense 27B to 31B at 8 bits, 117B to 125B MoE at 4 bits (Granite 4.2 30B, whose full-attention KV cache takes 8.6 GB per 32k session, holds eight sessions at 4 bits). With 72 GB, 3B-active MoE at 8 bits or BF16 and dense 27B to 31B at 4 or 8 bits. With 48 GB, 3B-active MoE at 4 or 8 bits (nine sessions at 8 bits for Nemotron 3.5 Lightning, Gemma 4 26B-A4B and gpt-oss-20b) and dense 27B to 30B at 4 bits; Gemma 4 31B, with its larger KV cache, holds seven 32k sessions at 4 bits there (size M). With 32 GB, 3B-active MoE at 4 bits (Laguna XS 2.1 and GLM-4.7-Flash hold only seven and six sessions: size M at most); a dense 27B fits at 4 bits but holds only six 32k sessions (size M). With 24 GB, gpt-oss-20b and Nemotron 3.5 Lightning hold nine sessions at 4 bits and Gemma 4 26B-A4B eight; the other 3B-active models hold two or fewer.

- **Volumes (estimate):**
  - 3B-active MoE models: on the RTX PRO 6000, within the targets up to L with both prompt cases (Qwen3.6-35B-A3B at 8 bits); on the RTX PRO 5000 (48 or 72 GB), up to L with typical prompts and M with documentation at the cap; on the GeForce RTX 5090, up to L with both (Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, gpt-oss-20b); on the R9700 and the Arc Pro B70, up to M with typical prompts and S with documentation at the cap, usable up to L (Gemma 4 26B-A4B: S with typical prompts; Laguna XS 2.1: usable up to M); on the W7900, whose prefill is estimated from declared compute only, up to S with typical prompts, usable up to M. The RTX PRO 4500 is not estimated (no compute figure published).
  - gpt-oss-120b and Mistral Small 4 at 4 bits on 96 GB: within the targets up to L with typical prompts and M with documentation at the cap. A dense 27B at 8 bits on 96 GB: within the targets up to M with typical prompts and S with documentation at the cap, usable up to L; on 32 to 72 GB cards a dense 27B is usable up to M.
  - 24 GB cards with Nemotron 3.5 Lightning at 4 bits: the RTX PRO 4000 within the targets up to M with typical prompts and S with documentation at the cap, the Arc Pro B60 up to S with typical prompts only; both usable up to L (with gpt-oss-20b the Arc Pro B60 is usable up to M).
  - Night capacity: 3,500 to 31,900 tests depending on card, model and concurrency.
- **Does not do:** 117B to 125B MoE below 96 GB; data-centre models; GeForce software "is not licensed for datacenter deployment" [hardware-43]; Radeon PRO and Radeon AI PRO cards are "not designed nor validated for datacenter usage" [hardware-54][hardware-55]. Card prices on 16 November 2026, host computer excluded: from EUR 1,659 (Arc Pro B70, 32 GB) to EUR 16,489.98 (RTX PRO 6000 Max-Q); 24 GB cards from EUR 889 (Arc Pro B60) (observed) [hardware-61][hardware-48][hardware-62].

6.3 Class 3: multi-GPU server

- **Hardware:** 4 x RTX PRO 6000 Server Edition (384 GB), 8-GPU nodes with H100, H200, B200 or MI355X (640 GB to 2.3 TB) (section 4.3). Single data-centre cards (L40S, H200 NVL) behave like class 2 for SparkTest.
- **Models:** data-centre models for customers that already run a model platform or want a larger model: Qwen3.5-397B-A17B (4 bits on 4 x RTX PRO 6000 Server, 8 bits on 8 x H100), GLM-5.3-Flash and DeepSeek-V4-Flash (8 bits on 4 x RTX PRO 6000 Server, weights only), Nemotron 3 Ultra, Mistral Large 3, DeepSeek-V4.1-Flash; GLM-5.3 and Kimi K3.1 Pro under custom licences (Kimi K3.1 Pro fits only the 8 x MI355X configuration at 4 bits among those compared); Qwen3.8-Flash-Next only after a legal opinion on its licence (table 3.3).
- **Volumes:** not a constraint; measured throughput is orders of magnitude above SparkTest's load at size L (section 5.5). One server can serve several SparkNodes, provided the endpoint is reachable only from inside the customer's network (section 7.4).
- **Does not do:** it is not an appliance; 8-GPU nodes cost USD 195,000 (used H100) to USD 470,000 (new B200) in the US in October 2026, and a DGX B200 or DGX B300 system draws up to about 14 kW (declared) [hardware-87][hardware-88][hardware-68][hardware-69]; their models' independent advantage on SparkTest's task is unmeasured; the KV cache of some 2026 architectures could not be derived (GLM-5.3-Flash, DeepSeek-V4-Flash).

6.4 Choice by size

The same estimates read by size. "3B-active MoE" means Qwen3.6-35B-A3B, Nemotron 3.5 Lightning, Laguna XS 2.1, Gemma 4 26B-A4B or gpt-oss-20b; exceptions are named.

SIZE	WITHIN THE TARGETS WITH BOTH PROMPT CASES	WITHIN THE TARGETS WITH TYPICAL PROMPTS ONLY	WITHIN THE HARD LIMITS ONLY (WORKS, MORE SLOWLY)
S (2 parallel runs)	GB10 with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1 at 4 bits; Radeon AI PRO R9700 and Arc Pro B70 with 3B-active MoE at 4 bits; RTX PRO 4000 with Nemotron 3.5 Lightning or gpt-oss-20b; RTX PRO 5000, RTX PRO 6000 and GeForce RTX 5090 with 3B-active MoE	GB10 with Gemma 4 26B-A4B or gpt-oss-20b; Arc Pro B60 with Nemotron 3.5 Lightning	Ryzen AI Max+ 395 or PRO 495 with 3B-active MoE at 4 bits
M (4 parallel runs)	RTX PRO 5000 (48 or 72 GB), RTX PRO 6000 or GeForce RTX 5090 with 3B-active MoE	GB10 with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1 (small margin, section 6.1); Radeon AI PRO R9700 and Arc Pro B70 with 3B-active MoE except Gemma 4 26B-A4B; RTX PRO 4000 with Nemotron 3.5 Lightning or gpt-oss-20b	GB10 with Gemma 4 26B-A4B, gpt-oss-20b, gpt-oss-120b or Mistral Small 4; Ryzen AI Max+ 395 or PRO 495 with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or Laguna XS 2.1; R9700 and Arc Pro B70 with Gemma 4 26B-A4B; Arc Pro B60 with Nemotron 3.5 Lightning or gpt-oss-20b
L (8 parallel runs)	RTX PRO 6000 with 3B-active MoE at 8 bits; GeForce RTX 5090 with Qwen3.6-35B-A3B, Nemotron 3.5 Lightning or gpt-oss-20b	RTX PRO 5000 (48 or 72 GB) with 3B-active MoE; RTX PRO 6000 with gpt-oss-120b or Mistral Small 4; GeForce RTX 5090 with Gemma 4 26B-A4B	GB10 with 3B-active MoE at 4 bits; Radeon AI PRO R9700 and Arc Pro B70 with 3B-active MoE except Laguna XS 2.1; RTX PRO 4000 with Nemotron 3.5 Lightning or gpt-oss-20b; Arc Pro B60 with Nemotron 3.5 Lightning; RTX PRO 6000 with a dense 27B at 8 bits

All entries are estimates (section 6) and must be confirmed with the qualification run of section 7.5; Ryzen AI Max+ systems are not usable at size L. The GeForce RTX 5090 carries the licence restriction of section 4.2; the Radeon PRO W7900 is covered in section 6.2.

7. Bringing your own model

7.1 Requirements

REQUIREMENT	WHAT IS NEEDED	WHY	SOURCES
Endpoint	OpenAI-compatible `POST {base}/v1/chat/completions`, non-streaming, reachable from the SparkNode inside the customer network	the agent sends one request per operation	[workload-serving-1]
Structured output	`response_format` of type `json_schema`, enforced by the engine, with the keywords of section 1.5	codify, repair and API insights parse the answer as JSON	[workload-serving-1] [workload-serving-2]
Context	32k tokens per request minimum, 64k recommended; sized per concurrent slot where the server shares one pool	prompts of 4k to 11k tokens, theoretical maximum above 32k	[workload-serving-10] [workload-serving-29]
Answer budget	`LLM_MAX_TOKENS` raised to 4,096	six long API tips need about 2,460 tokens	[workload-serving-10]
Reasoning	off for SparkTest calls (`reasoning_effort: "none"`, a template switch, a server flag, or a non-reasoning model); the agent asks for it (7.2), the kit shows whether the server complies (5.7)	reasoning tokens consume the budget and the timeouts	[workload-serving-16] [workload-serving-24] [workload-serving-29] [workload-serving-67]
Concurrency	as many simultaneous requests as the size's parallel runs plus one (3, 5 or 9), without queueing beyond the timeouts	sections 1.4 and 5.4	[workload-serving-4] [workload-serving-9]
Latency	p50 and p95 per operation within the targets of section 1.4 under that concurrency	the agent's timers include the queue	[workload-serving-10]
Data perimeter	endpoint inside the customer perimeter, bound to an internal address, with a key and TLS	prompts carry test descriptions, documentation, page text and labels, error messages and API URLs (secrets masked)	[workload-serving-1] [workload-serving-3]

7.2 Connecting the endpoint

The agent switches to the customer endpoint when `LLM_URL` is set. Settings read from the code [workload-serving-1]:

VARIABLE	MEANING	DEFAULT
`LLM_URL`	base URL; the agent appends `/v1/chat/completions`	not set
`LLM_MODEL`	served model name (vLLM `--served-model-name`, Ollama tag)	`default_model`
`LLM_API_KEY`	sent as `Authorization: Bearer` when set	empty
`LLM_MAX_TOKENS`	answer budget	2,048 (raise to 4,096)
`LLM_TEMPERATURE`	sampling temperature (seed 0 is always sent)	0.000001
`LLM_TIMEOUT_MS`	timeout of codify and repair calls	60,000
`LLM_EFFORT`	with `low` the request asks for no reasoning: `enable_thinking: false` at the top level and, since 17 November 2026, in `chat_template_kwargs`, plus `reasoning_effort: "low"`	`low`

The request carries `model`, one user message, `max_tokens`, `temperature`, `seed`, the reasoning fields above and, for codify, repair and API insights, `response_format: {"type": "json_schema", "json_schema": {"name": "sparktest_answer", "schema": ...}}` (the name since 17 November 2026); the answer is read from `choices[0].message.content` [workload-serving-1][workload-serving-67]. For an internal certificate authority, Node.js reads `NODE_EXTRA_CA_CERTS` or `NODE_USE_SYSTEM_CA=1` [workload-serving-65].

7.3 Serving software: compatibility on 16 and 17 November 2026

All cells are producer statements (documentation, release notes, source code) read by the workload study on 16 November 2026, applied to the agent's request as it is since 17 November 2026; so far only llama.cpp has served SparkTest's requests (section 5.7).

SERVER (VERSION CHECKED)	LICENCE	ACCEPTS THE AGENT'S REQUEST (SINCE 17 NOVEMBER 2026)	REASONING OFF	CONCURRENCY	NOTES	SOURCES
vLLM 0.33.0	Apache-2.0	yes, by its source code: the agent now sends `json_schema.name`, which it requires (until 16 November: no)	`reasoning_effort: "none"` or `chat_template_kwargs` (the agent sends the second); the top-level `enable_thinking` is ignored	continuous batching	usage statistics on by default (opt out with `VLLM_NO_USAGE_STATS` or `DO_NOT_TRACK`); release images built for CUDA 13.0, so NVIDIA driver 580 or later	[workload-serving-16][workload-serving-20][workload-serving-12] [workload-serving-22]
SGLang 0.5.23	Apache-2.0	yes, by its source code (until 16 November: no, `name` required)	`reasoning_effort: "none"`, which sets the template switch the agent now sends in `chat_template_kwargs`	not checked	RadixAttention prefix cache on by default	[workload-serving-24][workload-serving-26]
llama.cpp server b12445	MIT	yes (only `json_schema.schema` is read; measured with b12359, section 5.7)	`chat_template_kwargs` (sent by the agent; measured on Qwen3.6), `reasoning_effort: "none"` or `--reasoning-budget 0`	parallel slots, continuous batching	schema converted to a grammar; unsupported keywords skipped silently; `--offline`; binds 127.0.0.1 by default	[workload-serving-29][workload-serving-30][workload-serving-31]
Ollama 0.40.16	MIT	yes	`reasoning_effort: "none"`; the agent sends `"low"`, so use a model without reasoning or check with the kit that none comes back	one request per model by default (`OLLAMA_NUM_PARALLEL`)	4k default context below 24 GiB of GPU memory; no authentication on the local API	[workload-serving-33][workload-serving-35][workload-serving-36] [workload-serving-37][workload-serving-40]
LM Studio 0.4.x	app under its own terms of use (free for work); CLI and SDK MIT	not documented	not documented	continuous batching in the llama.cpp engine	serving works offline	[workload-serving-41][workload-serving-43][workload-serving-45] [workload-serving-46]
NVIDIA NIM 2.1.6	NVIDIA AI Enterprise in production: list USD 4,500 per GPU per year	follows its engine (vLLM 0.32.0 or SGLang 0.5.22), which requires `name`: sent since 17 November 2026	as its engine	as its engine	documented air-gap deployment	[workload-serving-50][workload-serving-52][workload-serving-54]
OpenVINO Model Server 2026.4.1	Apache-2.0	path is `/v3/chat/completions`	not checked	not checked	runs without the format, without error, when it cannot process it	[workload-serving-57]
Hugging Face TGI 3.3.7	Apache-2.0	no: different `response_format` shape	not checked	not checked	repository archived; maintenance mode	[workload-serving-48][workload-serving-49]

The changes that make the agent's request portable are small, and they are changes to the SparkTest agent, not to the customer's endpoint. Two were made on 17 November 2026: the request sends `json_schema.name`, and it asks for no reasoning with `chat_template_kwargs` and `reasoning_effort: "low"` besides the top-level flag (`"low"` is the lowest level of models such as gpt-oss, which always reason; `"none"` is not sent) [workload-serving-67]. Still open: `strict`, a configurable endpoint path, a default answer budget of 4,096 tokens and the recording of the `usage` token counts [workload-serving-16] [workload-serving-24][workload-serving-57]. Until the path is configurable, OpenVINO Model Server needs a rewrite in a reverse proxy (section 7.4).

Schema enforcement engines behind these servers (XGrammar, llguidance, Outlines, llama.cpp grammars) were compared independently on JSONSchemaBench, with small Llama models (1B to 8B parameters): on simple function-calling schemas their coverage was 0.93 to 0.96, on hard schemas collected from GitHub 0.03 to 0.41 [workload-serving-58]. SparkTest's three schemas are small (494 to 945 characters) and use common keywords, so coverage risk is low; compliance still has to be proven per model [workload-serving-17][workload-serving-30].

7.4 Checklist for the customer's IT team

01 Endpoint: `POST {base}/v1/chat/completions`, non-streaming, reachable from the SparkNode only; other paths (`/v3`) need a rewrite in a reverse proxy [workload-serving-57].

02 `LLM_MODEL` equal to the served model name [workload-serving-19].

03 Schema enforcement on, with an engine that covers the keywords of section 1.5; prove it is enforced, not just accepted, with a deliberately strict probe (an enum, a bounded array), because some servers skip unsupported keywords or formats silently [workload-serving-30][workload-serving-57].

04 Compact JSON where the server allows it (vLLM `disable_any_whitespace`), to save answer tokens [workload-serving-17].

05 Context of 32k per request, sized per slot where the pool is shared; for Ollama set `OLLAMA_CONTEXT_LENGTH` and `OLLAMA_NUM_PARALLEL` [workload-serving-29][workload-serving-35] [workload-serving-36].

06 Reasoning off for SparkTest calls (7.3), confirmed with the qualification kit, which reports any reasoning the server returns (section 5.7).

07 Concurrency of the size without queueing beyond the timeouts; measure p50 and p95 per operation.

08 Key and TLS on the endpoint (`--api-key` and TLS flags exist in vLLM, SGLang and llama.cpp; Ollama's local API needs a reverse proxy) [workload-serving-19][workload-serving-26][workload-serving-29] [workload-serving-37].

09 No outbound calls from the serving stack: `VLLM_NO_USAGE_STATS=1` or `DO_NOT_TRACK=1`, `HF_HUB_OFFLINE=1`, llama.cpp `--offline`, NIM in air-gap mode [workload-serving-20][workload-serving-61][workload-serving-29][workload-serving-52].

10 Request logging of the serving software reviewed, since prompts carry the data listed in 7.1.

11 Licence of the serving software and of the model checked (sections 3.3 and 7.3).

12 Qualification run passed (7.5).

7.5 Qualification with a test set

Public benchmarks do not measure SparkTest's task (section 8.1), so every model and endpoint is qualified on SparkTest's own cases before use, with the agent's own validation and scenario check active as in production [workload-serving-1][workload-serving-2]. SparkTech qualifies each SparkNode in the same way before use.

01 **Test set:** maintained tests written in Italian and in English on reference applications, including prompts with documentation at the cap, recorded interface changes for repairs, and seeded real defects (a repair must not turn a real defect into a pass).

02 **Runs:** codify every test from its description; run the suite on a changed version of the application; repeat at the size's concurrency.

03 **Measures:** JSON validity; share of tests codified at the first attempt; share of breakages repaired and accepted by the scenario check; repairs that hid a real defect; diagnosis written in the test's language; p50 and p95 time per operation; prompt and answer tokens from the `usage` field.

04 **Comparison:** the same run on a reference configuration. Acceptance thresholds are agreed with the customer; time thresholds follow section 1.4.

SparkTech's qualification kit (`agent/bin/qualify.js`) covers the first measures of this list on eight fixed tasks built with the agent's own prompts and checks, against any OpenAI-compatible endpoint, one request at a time; its first run, on two open-weight models, is section 5.7. Concurrency and the customer's own tests are measured on the node.

8. Risks and limits

8.1 Declared benchmarks and who measured them

- Vendor and independent figures for the same model differ widely: gpt-oss-120b scores 62.4 on SWE-bench Verified in its card (high effort, 477-task subset) against 33.60% at Vals AI and 26.0 on the bash-only board [benchmarks-49][benchmarks-18][benchmarks-23].
- Two competitors' runs of the same model differ by 11.7 points (Qwen3.6-27B on OSWorld-Verified) and 14.6 points (Gemma 4 31B on SWE-bench Verified); two harnesses on the same benchmark by 11 to 13 points [benchmarks-42][benchmarks-48][benchmarks-45][benchmarks-18][benchmarks-54].
- A model that does not follow a harness's protocol can score near zero: Nemotron 3.5 Super VL scores 0.00% on Terminal-Bench 4.0 with a minimal harness while its developer reports 60.67 on Terminal-Bench 2.1 with another harness [benchmarks-17][benchmarks-54].
- Independent data on node-scale models are thin: the Berkeley Function Calling Leaderboard V4 (last updated 2026-04-12) contains no 2026 open-weight release [benchmarks-32]; no independent Terminal-Bench 4.x, MCP Atlas or tau3 figure was found for Qwen3.8-27B, Gemma 4 31B, Muse Glimmer 30B, Qwen3.6-35B-A3B or Nemotron 3.5 Lightning.

For your IT team: do not choose a model from public leaderboards; the qualification of section 7.5 is the test that counts.

8.2 Saturation and contamination

On SWE-bench Verified, open-weight models released from July 2026 score 86.0% (Qwen3.8-27B) to 96.4% (DeepSeek-V4-Pro-0813) at Vals AI, which no longer runs it on new releases; OpenAI stopped reporting it in February 2026 after finding flawed tests in 59.4% of 138 audited tasks and signs that models had seen task details in training [benchmarks-18][benchmarks-24][benchmarks-25]. LiveCodeBench v6 problems predate the training data of 2026 models [benchmarks-54]. Scores near 90% on these sets say little about differences that matter for SparkTest.

8.3 Quality loss from quantization

- An independent academic study found that 8-bit quantization preserves accuracy (about 0.8% drop) while 4-bit methods lose up to 59% on long-context tasks, with losses that depend on method, model and task and grow when the input is not in English [method-4]. SparkTest's prompts are long and often in Italian.
- An independent single-model test found Qwen3.8-27B in Q4_K_M equal to BF16 on GPQA Diamond, IFBench and Terminal-Bench 2.1 [method-5]; Artificial Analysis measured the same index score for the BF16 and NVFP4 weights of Nemotron 3.5 Lightning [benchmarks-10]; NVIDIA reports 99.62% median accuracy recovery for NVFP4 weights of Nemotron 3.5 Super (vendor figure) [benchmarks-54] and 1% or less degradation from FP8 to NVFP4 on DeepSeek-R1-0528 (vendor figure) [method-2].
- Reading: 8-bit weights are the safer default where they fit; 4-bit weights must pass the qualification of section 7.5 on Italian and long prompts. The calculator gives speed at both precisions, not quality. On a compact system the sizes of section 6.1 need 4-bit weights.

8.4 Licences

The licensee question on a customer-sited node (SparkTech or the customer) decides whether revenue caps (Mistral Modified MIT: USD 20 million of monthly revenue), the Qwen Community License "AI Work Assistant" clause and the Step 5 hardware-device clause apply; it needs a lawyer [models-59][models-22][models-153]. Licence classifications differ between Hugging Face tags, Arena and Epoch AI for some models (Gemma 4 31B: Apache-2.0 in the card, "restricted use" at Epoch AI) [benchmarks-56][benchmarks-14]; the licence file of each repository is the reference.

8.5 Prices and availability

Prices rose 1.6 to 3.3 times between 2025 and November 2026 for several items (DGX Spark US list +74%, Framework Desktop 128 GB list +73%, Jetson AGX Thor developer kit +57%, RTX 5090 street 3.3 times its EU list price) [hardware-3][hardware-4][hardware-21][hardware-26][hardware-44][hardware-45]. The same chip sells at very different prices: on 16 November 2026 a GB10 system from ASUS with 1 TB of storage started at EUR 5,999 and NVIDIA's own, with 4 TB, at EUR 11,383.97 in Germany, while the 64 GB ASUS system (2 offers) cost more than the 128 GB one [hardware-9][hardware-8][hardware-10]. Purchase prices of AMD Instinct accelerators were not found. Any offer needs a dated quote.

8.6 Limits of the estimates and of the measurements

- **Calibration is in-sample and sparse.** Efficiency factors and effective compute come from 64 single-request decode points, 56 prefill points and 13 concurrency series, many from single community operators; only MLPerf (measured by the submitter and peer-reviewed; the DGX Spark and Ryzen AI Max+ 395 entries were submitted by the maker of the inference engine used, the Jetson AGX Thor entry by NVIDIA) and Artificial Analysis (independent) are controlled tests [measurements-26][measurements-24].
- **Spread.** Measured decode efficiency ranges from 0.38 to 0.91 on compact systems and from 0.21 to 0.82 on cards; effective prefill compute on GB10 ranges from 11.9 to 32.4 TFLOPS across models and builds (`out/tabelle.md`, table 1). Per-pair estimates can be off by tens of percent; the per-model deviation for Qwen3.6-35B-A3B on Ryzen AI Max+ 395 is 21% to 46% (section 5.3).
- **gpt-oss weighs on the class 1 calibration.** 16 of the 34 MoE decode points on unified memory are gpt-oss runs (11 of 14 on GB10); their median efficiency is 0.59 against 0.52 for the other models, and no measurement of Qwen3.6-35B-A3B on GB10 was found. Table 12 of `out/tabelle.md` gives the class 1 sizes at 0.52, 0.46 and 0.38: size M within the targets on GB10 is the verdict that moves.
- **What the model leaves out:** speculative decoding (which raised decode from 30.3 to 54.9 tok/s for Gemma 4 26B-A4B on DGX Spark [measurements-12]), prefix caching (which would shorten the repeated documentation of a codify session), the fixed state of linear-attention layers, the depth effect on prefill, the bandwidth used by the browser on unified-memory systems, tokenizer differences, and engines other than those measured. Prefill estimates for cards without measurements (RTX PRO 5000, RTX PRO 4000, W7900, Arc Pro B60, L40S, H200 NVL) are scaled from declared compute; RTX PRO 4500 has no declared figure to scale from and is not estimated.
- **Workload inputs:** 3.5 characters per token for Italian is an assumption; the night shares of breakages, flaky runs, defects and edits are assumptions (only the 3% planning figure and Google's 1.5% have a source) [workload-serving-9][workload-serving-63]; real token counts were measured on two models only (section 5.7: 3.4 to 4.3 characters per token for codify, repair and diagnosis prompts, 2.6 to 2.8 for the API statistics prompt, and API analyses of 700 to 881 answer tokens against the 418 assumed in section 1.2) [workload-serving-10].
- **Bursts are an upper estimate:** the calculator assumes that all concurrent calls arrive together, and the documentation-at-cap case assumes that every concurrent codify round carries the full 24,000 characters.
- **Live pages** (InferenceX, Artificial Analysis) may change after 16 November 2026 [measurements-30] [measurements-25].

For your IT team: treat the sizes that sit near a limit in section 6 (GB10 at size M within the targets, Ryzen AI Max+ 395 at any size) as provisional until measured on your hardware (section 7.5).

8.7 Italian

No Italian-specific evaluation of the 2026 open-weight models was found. Multilingual aggregates show a small gap (Artificial Analysis Italian index: Qwen3.5-397B-A17B 93 against 95 for the first model) [benchmarks-13]. The one 2026 study on ITALIC (10,000 questions from Italian public exams) covers models up to 20B: Ministral-3-8B 80.2% to 87.4%, Gemma 3 12B 76.5% to 79.0%, Qwen3-8B 70.8% to 77.4%, gpt-oss-20b 26.1% to 76.9% depending on the prompt format, against 86.1% to 86.8% for GPT-5 nano [benchmarks-57]. Italian appears in the cards of Mistral, Nemotron, Granite and EuroLLM models; Qwen, Gemma and Muse claim broad coverage without naming it (table 3.1). The qualification set of section 7.5 must therefore include Italian tests.

9. Method and reproducibility

9.1 Sources and labels

Five research notes, compiled on 16 November 2026 from primary sources where possible (model cards and configuration files, vendor specifications, raw benchmark files, source code of the serving software), are the inputs: `fonti/models.*` (120 models), `fonti/hardware.*` (57 product, system and price records), `fonti/measurements.*` (185 measurement rows), `fonti/benchmarks.*` (607 benchmark results, 610 after the verification) and `fonti/workload-serving.*` (SparkTest workload from the code, serving software). No model weights were downloaded and no model was run for the estimates of this study; the measurement of

section 5.7 was made afterwards, on 17 November 2026, and is labelled as measured by SparkTech. Five references were added by the author for the method ([method-1] to [method-5]). Each note ends with the corrections made after the verification of 17 November 2026.

9.2 Formulas

QUANTITY	FORMULA	LABEL
Weights	parameters held in memory x bits per weight / 8	estimate
KV cache, grouped-query attention	2 (K and V) x cache-holding layers x KV heads x head dimension x bytes per value x tokens	estimate (configuration values declared)
KV cache, multi-head latent attention	cache-holding layers x (latent + RoPE dimensions) x bytes per value x tokens	estimate
Sliding-window layers	as above, with tokens capped at the window	estimate
Fit	weights + sessions x KV(32k) + reserve <= memory	estimate
Decode bound	bandwidth / (active parameters x bits / 8 + KV read at the current depth)	estimate
Decode estimate	decode bound x efficiency (section 9.3)	estimate
Prefill estimate	effective compute / (2 x active parameters)	estimate
Time of one call	prompt tokens / prefill + answer tokens / decode, with decode at the depth prompt + half the answer	estimate
Burst of c calls	c x prompt tokens / prefill + answer tokens / (decode x c^(beta - 1))	estimate
Night capacity	night window / (prompt tokens per test / prefill + answer tokens per test / (decode x c^beta)), with decode at a depth of 5,000 tokens	estimate
Size verdict	largest size whose bursts of codify, repair and diagnosis all meet the targets (or the hard limits) and whose concurrent 32k sessions fit; computed with typical prompts and with the codify prompt at the documentation cap	estimate
Usable	the smaller of the two hard-limit verdicts	estimate

9.3 Calibration values (computed by the calculator from `dati/measurements.json`)

FACTOR	VALUE	BASIS
Decode efficiency, unified memory, dense	0.84 (0.67 to 0.91; 11 rows)	median of measured / bound, no speculative decoding, depth up to 4,096
Decode efficiency, unified memory, MoE	0.56 (0.38 to 0.78; 34 rows); 0.52 without the 16 gpt-oss rows (sensitivity only)	same
Decode efficiency, GPU card, dense	0.67 (0.37 to 0.82; 7 rows)	same
Decode efficiency, GPU card, MoE	0.28 (0.21 to 0.39; 12 rows)	same
Effective prefill compute, GB10	19.7 TFLOPS MoE (11 rows); 38.9 dense (2 rows)	median of measured prefill x 2 x active parameters
Effective prefill compute, Ryzen AI Max+ 395	6.7 TFLOPS MoE (21 rows); 20.2 dense (7 rows)	same; the PRO 495 uses these x 61/60 (declared GPU FP16 ratio)
Effective prefill compute, cards	RTX PRO 6000 69.6 MoE; RTX 5090 56.5 MoE; R9700 22.8 MoE, 58.4 dense; Arc Pro B70 69.7 dense	same
Dense / MoE ratio	2.57	platforms measured with both
kappa (effective / declared dense 8-bit compute)	0.173 dense, 0.067 MoE (5 cards)	used for cards without prefill measurements
Concurrency exponent beta	0.44 MoE (9 series); 0.96 dense (4 series)	measured concurrency series

One community run on a Ryzen AI Max+ PRO 495 system gives 16.1 TFLOPS of effective prefill compute for GLM-5.3-Flash, 2.4 times the Ryzen AI Max+ 395 median [measurements-21]; with a single row and an unusual architecture it is not used. The two Ryzen AI Max+ 395 backends added after the verification (row T1-41) are kept for comparison and not used, so as not to add more gpt-oss points to the calibration.

9.4 Assumptions

ASSUMPTION	VALUE	BASIS
Bits per weight	BF16 16; 8-bit 8.5; 4-bit 4.5	llama.cpp Q8_0 8.5008 [method-1]; NVFP4 4.5 [method-2]; public GGUF 4-bit files measured 4.3 to 5.4 bits per weight (file size / parameters, [measurements-20])
KV cache precision	BF16	FP8 halves the KV figures
Session length for the fit	32,768 tokens	minimum context recommended by the workload study [workload-serving-10]
Reserve	16 GB on unified memory; 10% of memory on a GPU	assumption of this study; for scale, vLLM uses 92% of GPU memory by default when its KV cache is not extensible [method-3]
Concurrency	3, 5, 9 requests for S, M, L	[workload-serving-4][workload-serving-9]
Night	6 hours; release-night load, high bound per test (3,394 prompt and 233 answer tokens, with the rule of 17 November 2026; 4,975 and 256 before)	[workload-serving-10][workload-serving-67]
Memory units	weights and KV in GB of 10^9 bytes; capacities as declared	where declared GB are binary, usable bytes are up to 7% larger

9.5 Files and how to rerun

```
marketing/research/
  fonti/              research notes (Markdown) and their data (JSON), 16 November 2026, with
the corrections of 17 November 2026
  dati/normalizza.mjs  builds dati/*.json from fonti/ (curated choices written in the script)
  dati/models.json     120 models, normalized, with sources
  dati/hardware.json   39 platforms, 18 systems, 22 configurations compared, assumptions,
sources
  dati/measurements.json  185 measurement rows mapped to models and hardware, quantization
rates, sources
  dati/workload.json   SparkTest sizes, operations, night loads, serving records, sources
  dati/sources.json    all 455 sources
  calcolatore.mjs      the calculator (Node, no dependencies)
  out/tabelle.md       calculator output, 12 tables
  out/tabelle.json     the same, machine-readable
  qualifica/           the measurement of section 5.7: kit reports, scripts used on the test
server, data
```

Commands, from `marketing/research/`: `node dati/normalizza.mjs` then `node calcolatore.mjs`. Both run in well under a second, without network access. Every number in sections 5 and 6 can be traced to `out/tabelle.json`; every input to a source id in `dati/*.json`.

References

All sources were accessed on 16 November 2026 unless the entry says otherwise. Entries marked internal are SparkTech's own repository and have no public URL. Entries marked "rechecked 2026-11-17" were opened again during the verification.

SparkTest workload and serving software (note `fonti/workload-serving.md`)

- [workload-serving-1] SparkTest agent source, `agent/src/llm.js` (prompt builders, HTTP client, timeouts, insights schema). SparkTech, internal repository, working tree of 2026-11-16 22:09. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-2] SparkTest agent source, `agent/src/plan.js` (plan DSL, `PLAN_SCHEMA`, validation). SparkTech, internal, working tree of 2026-11-16 22:02. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-3] SparkTest agent source, `agent/src/harness.js` (page snapshot, network capture caps). SparkTech, internal, working tree of 2026-11-16 22:29. Accessed 2026-11-16. Internal, no public URL.

- [workload-serving-4] SparkTest agent source, `agent/src/codify.js`, `agent/src/runner.js`, `agent/src/config.js` (codification budgets, repair attempts, diagnosis, rebuild after failure until 2026-11-16, parallel slots). SparkTech, internal, working tree of 2026-11-16 22:31. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-5] SparkTest console source, `web/lib/suggest-tests.ts`, `web/lib/chat-prompt.ts`, `web/lib/prompt-data.ts`, `web/app/api/portal/chat/route.ts`. SparkTech, internal, working tree of 2026-11-16 22:40. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-6] SparkTest console source, `web/lib/schedule.ts` (nightly suite), `web/lib/api-insights.ts` (on-request analysis and caps), `web/app/api/agent/claim/route.ts` (documentation sent to the agent). SparkTech, internal. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-9] SparkTest product notes, `docs/VISION.md` (on-premise sizes S, M, L: 100, 400, 1,500 tests; 2, 4, 8 parallel runs) and `docs/BUSINESS-MODEL.md` (3% repair rate used for planning). SparkTech, internal. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-10] This study's measurement scripts and raw outputs: `marketing/research/dati/workload-serving-prompt-sizes.mjs` (+ `.out.txt`) and `marketing/research/dati/workload-serving-night-load.mjs` (+ `.out.txt`). 2026-11-16. Accessed 2026-11-16. Internal, no public URL.
- [workload-serving-11] "What are tokens and how to count them?", OpenAI Help Center, https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them. Direct fetch refused (HTTP 403); rule of thumb read in the search engine's rendering of the page. Accessed 2026-11-16.
- [workload-serving-12] vLLM v0.33.0 release notes, vLLM project, https://github.com/vllm-project/vllm/releases/tag/v0.33.0 (notes published with v0.33.0rc3 on 2026-11-02). Accessed 2026-11-16.
- [workload-serving-16] vLLM source code at tag v0.33.0: `vllm/entrypoints/generate/base/protocol.py` (`JsonSchemaResponseFormat`), `vllm/entrypoints/serve/engine/protocol.py` (`OpenAIBaseModel`, extra fields), `vllm/entrypoints/openai/chat_completion/protocol.py` (`reasoning_effort`), https://github.com/vllm-project/vllm/tree/v0.33.0/vllm/entrypoints. Accessed 2026-11-16.
- [workload-serving-17] vLLM source code at tag v0.33.0: `vllm/config/structured_outputs.py` and `vllm/v1/structured_output/backend_xgrammar.py`, https://github.com/vllm-project/vllm/tree/v0.33.0/vllm. Accessed 2026-11-16.
- [workload-serving-19] "vllm serve" CLI reference, vLLM documentation, https://docs.vllm.ai/en/latest/cli/serve.html. Accessed 2026-11-16.
- [workload-serving-20] "Usage Stats Collection", vLLM documentation, https://docs.vllm.ai/en/latest/usage/usage_stats.html. Accessed 2026-11-16.
- [workload-serving-22] "CUDA Toolkit Release Notes" (toolkit and minimum driver versions), NVIDIA, https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html. Accessed 2026-11-16.
- [workload-serving-24] SGLang source code at tag v0.5.23: `python/sglang/srt/entrypoints/openai/protocol.py`, https://github.com/sgl-project/sglang/blob/v0.5.23/python/sglang/srt/entrypoints/openai/protocol.py. Accessed 2026-11-16.
- [workload-serving-26] "Server Arguments", SGLang documentation, https://docs.sglang.io/docs/advanced_features/server_arguments. Accessed 2026-11-16.
- [workload-serving-29] "llama-server" README, ggml-org/llama.cpp, https://github.com/ggml-org/llama.cpp/blob/master/tools/server/README.md. Accessed 2026-11-16.
- [workload-serving-30] "GBNF Guide", section "JSON Schemas → GBNF", ggml-org/llama.cpp, https://github.com/ggml-org/llama.cpp/blob/master/grammars/README.md. Accessed 2026-11-16.
- [workload-serving-31] llama.cpp source at b12445: `tools/server/server-common.cpp`, https://github.com/ggml-org/llama.cpp/blob/b12445/tools/server/server-common.cpp. Accessed 2026-11-16.
- [workload-serving-33] "OpenAI compatibility", Ollama documentation at v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/api/openai-compatibility.mdx. Accessed 2026-11-16.
- [workload-serving-35] "Context length", Ollama documentation at v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/context-length.mdx. Accessed 2026-11-16.
- [workload-serving-36] "FAQ" (parallel requests, queue), Ollama documentation at v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/faq.mdx. Accessed 2026-11-16.

- [workload-serving-37] "Authentication", Ollama documentation at v0.40.16, https://github.com/ollama/ollama/blob/v0.40.16/docs/api/authentication.mdx. Accessed 2026-11-16.
- [workload-serving-40] Ollama source at v0.40.16: `openai/openai.go`, https://github.com/ollama/ollama/blob/v0.40.16/openai/openai.go. Accessed 2026-11-16.
- [workload-serving-41] "Structured Output" (OpenAI compatibility), LM Studio documentation, https://lmstudio.ai/docs/developer/openai-compat/structured-output. Accessed 2026-11-16.
- [workload-serving-43] "LM Studio is free for use at work", LM Studio blog, https://lmstudio.ai/blog/free-for-work, 2025-07-08. Accessed 2026-11-16.
- [workload-serving-45] "Offline Operation", LM Studio documentation, https://lmstudio.ai/docs/app/offline. Accessed 2026-11-16.
- [workload-serving-46] LM Studio repositories `lmstudio-ai/lms` and `lmstudio-ai/lmstudio-js` (MIT licence); no public repository `lmstudio-ai/llmster` (GitHub API: not found), https://github.com/lmstudio-ai. Accessed 2026-11-16.
- [workload-serving-48] text-generation-inference README (maintenance mode notice), repository status (archived) and releases (v3.3.7, 2025-12-19), Hugging Face, https://github.com/huggingface/text-generation-inference. Accessed 2026-11-16.
- [workload-serving-49] TGI "Guidance" documentation and router source `router/src/lib.rs` (`GrammarType`, `JsonSchemaConfig`), Hugging Face, https://github.com/huggingface/text-generation-inference/blob/main/router/src/lib.rs. Accessed 2026-11-16.
- [workload-serving-50] "Release Notes", NVIDIA NIM for LLM and VLM (release 2.1.6), https://docs.nvidia.com/nim/large-language-models/latest/about-nim-llm/release-notes.html, page updated 2026-11-16. Accessed 2026-11-16.
- [workload-serving-52] "Air-Gap Deployment", NVIDIA NIM for LLM and VLM, https://docs.nvidia.com/nim/large-language-models/latest/deployment/air-gap-deployment.html. Accessed 2026-11-16.
- [workload-serving-54] "Pricing", NVIDIA AI Enterprise Licensing Guide, https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/pricing.html, updated 2026-09-02. Accessed 2026-11-16.
- [workload-serving-57] "OpenAI API chat/completions endpoint", OpenVINO Model Server documentation 2026, https://docs.openvino.ai/2026/model-server/ovms_docs_rest_api_chat.html; release 2026.4.1 (2026-10-06) and licence (Apache-2.0), https://github.com/openvinotoolkit/model_server. Accessed 2026-11-16.
- [workload-serving-58] S. Geng, H. Cooper, M. Moskal, S. Jenkins, J. Berman, N. Ranchin, R. West, E. Horvitz, H. Nori, "JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models", arXiv:2501.10868v3, https://arxiv.org/abs/2501.10868, 2025-02-27 (v1 2025-01-18). Accessed 2026-11-16.
- [workload-serving-61] "Environment variables" (`HF_HUB_OFFLINE`, `HF_HUB_DISABLE_TELEMETRY`, `DO_NOT_TRACK`), Hugging Face Hub documentation, https://huggingface.co/docs/huggingface_hub/package_reference/environment_variables. Accessed 2026-11-16.
- [workload-serving-63] J. Micco, "Flaky Tests at Google and How We Mitigate Them", Google Testing Blog, https://testing.googleblog.com/2016/05/flaky-tests-at-google-and-how-we.html, 2016-05-27. Accessed 2026-11-16.
- [workload-serving-65] "Command-line API" (`NODE_EXTRA_CA_CERTS`, `NODE_USE_SYSTEM_CA`), Node.js documentation, https://nodejs.org/api/cli.html. Accessed 2026-11-16.
- [workload-serving-67] SparkTest agent source after the changes of 2026-11-17: `agent/src/runner.js` (no new codification after a failed run; agent 0.2.1, commit of 2026-11-17 00:58), `agent/src/llm.js` (scenario check on expected values and typed data; request with `json_schema.name`, `chat_template_kwargs` and `reasoning_effort`, commit of 2026-11-17 01:40). SparkTech, internal. Accessed 2026-11-17. Internal, no public URL.

Models (note `fonti/models.md`)

- [models-2] Berkeley Function Calling Leaderboard (BFCL) V4, page and data_overall.csv. Gorilla project, UC Berkeley. https://gorilla.cs.berkeley.edu/leaderboard.html. page states Last Updated 2026-04-12; CSV: https://gorilla.cs.berkeley.edu/data_overall.csv. Accessed 2026-11-16.

- [models-3] Alibaba launches Qwen3.5 with open-weight 397B model and broader language support (secondary). AlternativeTo. https://alternativeto.net/news/2026/2/alibaba-launches-qwen3-5-with-open-weight-397b-model-and-broader-language-support. February 2026. Accessed 2026-11-16.
- [models-6] Qwen/Qwen3-Coder-Next: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3-Coder-Next. Hugging Face repo created 2026-01-30, last modified 2026-02-03. Accessed 2026-11-16.
- [models-10] Qwen/Qwen3.5-122B-A10B: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.5-122B-A10B. Hugging Face repo created 2026-02-24, last modified 2026-04-24. Accessed 2026-11-16.
- [models-12] Qwen/Qwen3.6-35B-A3B: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Hugging Face repo created 2026-04-15, last modified 2026-11-07. Accessed 2026-11-16.
- [models-14] Qwen3.6 35B A3B: model page. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-6-35b-a3b. no data date shown on page. Accessed 2026-11-16.
- [models-15] Qwen/Qwen3.8-27B: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.8-27B. Hugging Face repo created 2026-08-05, last modified 2026-11-07. Accessed 2026-11-16.
- [models-16] Qwen3.8 27B: model page. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-8-27b. no data date shown on page. Accessed 2026-11-16.
- [models-17] Open weights models comparison (Intelligence Index v5.0.1). Artificial Analysis. https://artificialanalysis.ai/models/open-source. no data date shown on page. Accessed 2026-11-16.
- [models-20] Qwen/Qwen3.5-397B-A17B: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Hugging Face repo created 2026-02-16, last modified 2026-04-24. Accessed 2026-11-16.
- [models-21] Qwen/Qwen3.8-Flash-Next: model card (README.md), config.json and Hub API metadata. Qwen (Hugging Face). https://huggingface.co/Qwen/Qwen3.8-Flash-Next. Hugging Face repo created 2026-08-24, last modified 2026-11-07. Accessed 2026-11-16.
- [models-22] Qwen Community License 1.0 (LICENSE of Qwen3.8-Flash-Next). Qwen. https://huggingface.co/Qwen/Qwen3.8-Flash-Next/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-24] Qwen3.8-Flash-Next: model page. Artificial Analysis. https://artificialanalysis.ai/models/qwen3-8-flash-next. no data date shown on page. Accessed 2026-11-16.
- [models-30] DeepSeek V4 Preview Release. DeepSeek API Docs, News. https://api-docs.deepseek.com/news/news260424/. 2026-04-24. Accessed 2026-11-16.
- [models-32] deepseek-ai/DeepSeek-V4-Flash-0731: model card (README.md), config.json and Hub API metadata. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731. Hugging Face repo created 2026-07-31, last modified 2026-08-01. Accessed 2026-11-16.
- [models-34] deepseek-ai/DeepSeek-V4.1-Flash: model card (README.md), config.json and Hub API metadata. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Hugging Face repo created 2026-09-10, last modified 2026-10-16. Accessed 2026-11-16.
- [models-35] deepseek-ai/DeepSeek-V4.1-Flash-1017: model card (README.md), config.json and Hub API metadata. deepseek-ai (Hugging Face). https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017. Hugging Face repo created 2026-10-17, last modified 2026-11-10. Accessed 2026-11-16.
- [models-36] openai/gpt-oss-20b: model card (README.md), config.json and Hub API metadata. openai (Hugging Face). https://huggingface.co/openai/gpt-oss-20b. Hugging Face repo created 2025-08-04, last modified 2025-08-26. Accessed 2026-11-16.
- [models-37] Introducing gpt-oss. OpenAI. https://openai.com/index/introducing-gpt-oss/. 2025-08-05 (date per search results; page not opened). Accessed 2026-11-16.
- [models-38] gpt-oss-20b: model page. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-20b. no data date shown on page. Accessed 2026-11-16.
- [models-39] openai/gpt-oss-120b: model card (README.md), config.json and Hub API metadata. openai (Hugging Face). https://huggingface.co/openai/gpt-oss-120b. Hugging Face repo created 2025-08-04, last modified 2025-08-26. Accessed 2026-11-16.
- [models-40] gpt-oss-120b (High): intelligence, performance and price analysis. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-120b. no data date shown on page. Accessed 2026-11-16.

- [models-44] Llama 4 Community License Agreement (LICENSE). Meta Platforms. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-45] Llama 4 Acceptable Use Policy (USE_POLICY.md). Meta, meta-llama/llama-models on GitHub. https://raw.githubusercontent.com/meta-llama/llama-models/main/models/llama4/USE_POLICY.md. Llama 4 release, April 2025. Accessed 2026-11-16.
- [models-48] meta-models/Muse-Glimmer-30B: model card (README.md), config.json and Hub API metadata. meta-models (Hugging Face). https://huggingface.co/meta-models/Muse-Glimmer-30B. Hugging Face repo created 2026-08-09, last modified 2026-10-22. Accessed 2026-11-16.
- [models-51] Muse Model License, version Muse Spark Open (LICENSE.md). Meta Platforms. https://huggingface.co/meta-models/Muse-Spark-Open-816B-A42B/blob/main/LICENSE.md. license text as published in the repository. Accessed 2026-11-16.
- [models-52] mistralai/Mistral-Small-3.2-24B-Instruct-2506: model card (README.md), config.json and Hub API metadata. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Hugging Face repo created 2025-06-19, last modified 2025-12-22. Accessed 2026-11-16.
- [models-55] mistralai/Ministral-3-14B-Instruct-2512: model card (README.md), config.json and Hub API metadata. mistralai (Hugging Face). https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512. Hugging Face repo created 2025-10-31, last modified 2026-07-15. Accessed 2026-11-16.
- [models-56] mistralai/Mistral-Small-4-119B-2603: model card (README.md), config.json and Hub API metadata. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Small-4-119B-2603. Hugging Face repo created 2026-01-23, last modified 2026-07-15. Accessed 2026-11-16.
- [models-57] Mistral Small 4: model page. Artificial Analysis. https://artificialanalysis.ai/models/mistral-small-4. no data date shown on page. Accessed 2026-11-16.
- [models-59] Modified MIT License (LICENSE of Mistral-Medium-3.5-128B). Mistral AI. https://huggingface.co/mistralai/Mistral-Medium-3.5-128B/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-61] Modified MIT License (LICENSE of Devstral-2-123B-Instruct-2512). Mistral AI. https://huggingface.co/mistralai/Devstral-2-123B-Instruct-2512/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-63] Mistral AI Non-Production License (MNPL-0.1). Mistral AI. https://mistral.ai/licences/MNPL-0.1.md. license text as published in the repository. Accessed 2026-11-16.
- [models-64] mistralai/Mistral-Large-3-675B-Instruct-2512: model card (README.md), config.json and Hub API metadata. mistralai (Hugging Face). https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512. Hugging Face repo created 2025-11-28, last modified 2026-07-15. Accessed 2026-11-16.
- [models-71] Gemma 4: Expanding the Gemmaverse with Apache 2.0. Google Open Source Blog. https://opensource.googleblog.com/2026/03/gemma-4-expanding-the-gemmaverse-with-apache-20.html. 2026-04-02. Accessed 2026-11-16.
- [models-72] google/gemma-4-31B-it: model card (README.md), config.json and Hub API metadata. google (Hugging Face). https://huggingface.co/google/gemma-4-31B-it. Hugging Face repo created 2026-03-11, last modified 2026-07-20. Accessed 2026-11-16.
- [models-73] Gemma 4 31B: model page. Artificial Analysis. https://artificialanalysis.ai/models/gemma-4-31b. no data date shown on page. Accessed 2026-11-16.
- [models-74] google/gemma-4-26B-A4B-it: model card (README.md), config.json and Hub API metadata. google (Hugging Face). https://huggingface.co/google/gemma-4-26B-A4B-it. Hugging Face repo created 2026-03-11, last modified 2026-07-20. Accessed 2026-11-16.
- [models-75] Gemma 4 26B A4B: model page. Artificial Analysis. https://artificialanalysis.ai/models/gemma-4-26b-a4b. no data date shown on page. Accessed 2026-11-16.
- [models-79] zai-org/GLM-4.7-Flash: model card (README.md), config.json and Hub API metadata. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-4.7-Flash. Hugging Face repo created 2026-01-19, last modified 2026-01-29. Accessed 2026-11-16.
- [models-82] zai-org/GLM-5: model card (README.md), config.json and Hub API metadata. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5. Hugging Face repo created 2026-02-11, last modified 2026-08-11. Accessed 2026-11-16; rechecked 2026-11-17 ("GLM-5 scales from 355B parameters (32B active) to 744B parameters (40B active)").

- [models-85] zai-org/GLM-5.3: model card (README.md), config.json and Hub API metadata. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5.3. Hugging Face repo created 2026-08-25, last modified 2026-10-14. Accessed 2026-11-16; rechecked 2026-11-17 (same base model as GLM-5.2; no parameter count in the card).
- [models-86] GLM-5.3 License (LICENSE). Z.ai. https://huggingface.co/zai-org/GLM-5.3/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-87] zai-org/GLM-5.3-Flash: model card (README.md), config.json and Hub API metadata. zai-org (Hugging Face). https://huggingface.co/zai-org/GLM-5.3-Flash. Hugging Face repo created 2026-08-25, last modified 2026-10-15. Accessed 2026-11-16.
- [models-88] GLM-5.3-Flash: model page. Artificial Analysis. https://artificialanalysis.ai/models/glm-5-3-flash. no data date shown on page. Accessed 2026-11-16.
- [models-100] moonshotai/Kimi-K3.1-Pro: model card (README.md), config.json and Hub API metadata. moonshotai (Hugging Face). https://huggingface.co/moonshotai/Kimi-K3.1-Pro. Hugging Face repo created 2026-11-02, last modified 2026-11-05. Accessed 2026-11-16.
- [models-101] Kimi K3.1-Pro License (LICENSE). Moonshot AI. https://huggingface.co/moonshotai/Kimi-K3.1-Pro/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-113] ibm-granite/granite-4.2-30b: model card (README.md), config.json and Hub API metadata. ibm-granite (Hugging Face). https://huggingface.co/ibm-granite/granite-4.2-30b. Hugging Face repo created 2026-08-07, last modified 2026-09-04. Accessed 2026-11-16.
- [models-115] NVIDIA Nemotron Open Model License. NVIDIA. https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-nemotron-open-model-license/. v. December 15, 2025. Accessed 2026-11-16.
- [models-118] nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16: model card (README.md), config.json and Hub API metadata. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16. Hugging Face repo created 2026-08-01, last modified 2026-08-24. Accessed 2026-11-16.
- [models-119] OpenMDW License Agreement 1.1 (LICENSE of NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16). NVIDIA. https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-121] nvidia/NVIDIA-Nemotron-3.5-Bolt-74B-A9B-BF16: model card (README.md), config.json and Hub API metadata. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Bolt-74B-A9B-BF16. Hugging Face repo created 2026-10-09, last modified 2026-10-23. Accessed 2026-11-16.
- [models-122] nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-BF16: model card (README.md), config.json and Hub API metadata. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-BF16. Hugging Face repo created 2026-10-12, last modified 2026-10-22. Accessed 2026-11-16.
- [models-123] nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16: model card (README.md), config.json and Hub API metadata. nvidia (Hugging Face). https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16. Hugging Face repo created 2026-06-03, last modified 2026-08-24. Accessed 2026-11-16.
- [models-124] poolside/Laguna-XS-2.1: model card (README.md), config.json and Hub API metadata. poolside (Hugging Face). https://huggingface.co/poolside/Laguna-XS-2.1. Hugging Face repo created 2026-06-20, last modified 2026-08-19. Accessed 2026-11-16.
- [models-132] utter-project/EuroLLM-22B-Instruct-2512: model card (README.md), config.json and Hub API metadata. utter-project (Hugging Face). https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512. Hugging Face repo created 2025-12-05, last modified 2026-07-31. Accessed 2026-11-16.
- [models-133] swiss-ai/Apertus-v1.5-8B: model card (README.md), config.json and Hub API metadata. swiss-ai (Hugging Face). https://huggingface.co/swiss-ai/Apertus-v1.5-8B. Hugging Face repo created 2026-07-24, last modified 2026-11-03. Accessed 2026-11-16.
- [models-137] AlmawaveLabs/Velvet2-3B: model card (README.md), config.json and Hub API metadata. AlmawaveLabs (Hugging Face). https://huggingface.co/AlmawaveLabs/Velvet2-3B. Hugging Face repo created 2026-10-15, last modified 2026-11-09. Accessed 2026-11-16.
- [models-146] MiniMax M2.7 Non-Commercial License (LICENSE). MiniMax. https://huggingface.co/MiniMaxAI/MiniMax-M2.7/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.

- [models-153] Step 5 Preview license (LICENSE.txt). StepFun. https://huggingface.co/stepfun-ai/Step-5-Preview/blob/main/LICENSE.txt. license text as published in the repository. Accessed 2026-11-16.
- [models-162] ERNIE 5.5 Flash Base Model License Agreement v1.0 (LICENSE). Baidu. https://huggingface.co/baidu/ERNIE-5.5-Flash-Base/blob/main/LICENSE. license text as published in the repository. Accessed 2026-11-16.
- [models-175] Structured Outputs (vLLM documentation, latest). vLLM project. https://docs.vllm.ai/en/latest/features/structured_outputs.html. undated developer docs. Accessed 2026-11-16.
- [models-176] llama.cpp HTTP server README. ggml-org. https://github.com/ggml-org/llama.cpp/blob/master/tools/server/README.md. undated, master branch. Accessed 2026-11-16.
- [models-177] Qwen4-Max-Preview (secondary). AI TL;DR. https://ai-tldr.dev/releases/qwen-qwen4-max-preview/. 2026-11-11. Accessed 2026-11-16.
- [models-178] Qwen4: four variants, no confirmed open weights (secondary). Tech Insider. https://tech-insider.org/qwen4-four-variants-no-confirmed-open-weights-alibaba-2026/. 2026 (Apsara Conference 2026-09-22 per article). Accessed 2026-11-16.

Hardware (note `fonti/hardware.md`)

- [hardware-1] NVIDIA. "NVIDIA DGX Spark" (product page, technical specifications) https://www.nvidia.com/en-us/products/workstations/dgx-spark/. Accessed 2026-11-16.
- [hardware-2] NVIDIA. "DGX Spark User Guide: Hardware Overview". https://docs.nvidia.com/dgx/dgx-spark/hardware.html. Accessed 2026-11-16.
- [hardware-3] NVIDIA Developer Forums. "2/23/2026 Price Change Announcement" (DGX Spark Founders Edition) https://forums.developer.nvidia.com/t/2-23-2026-price-change-announcement/361713. Accessed 2026-11-16.
- [hardware-4] The Register, Tobias Mann. "Nvidia debuts $4,999 DGX Spark with half the RAM and storage, amid memory crunch". https://www.theregister.com/a/5300622. Accessed 2026-11-16.
- [hardware-5] ServeTheHome, John Lee. "NVIDIA DGX Spark 64GB launched and big 128GB GB10 price increases". https://www.servethehome.com/nvidia-dgx-spark-64gb-launched-and-big-128gb-gb10-price-increases/. Accessed 2026-11-16.
- [hardware-6] NVIDIA. "NVIDIA DGX Spark Warranty". https://www.nvidia.com/en-us/support/warranty/dgx-spark/. Accessed 2026-11-16.
- [hardware-8] geizhals.de. "PNY NVIDIA DGX Spark Founders Edition (DGXSPARK-FOUNEDIT-EU)". https://geizhals.de/pny-nvidia-dgx-spark-dgxspark-founedit-eu-a3552134.html. Accessed 2026-11-16.
- [hardware-9] geizhals.de. "ASUS Ascent (DGX Spark) GX10-GG0003BN, 128 GB, 1 TB". https://geizhals.de/asus-ascent-gx10-gg0003bn-90ms0371-m00030-a3551306.html. Accessed 2026-11-16.
- [hardware-10] geizhals.de. "ASUS Ascent (DGX Spark)" variant overview (incl. GX10-GG0062BN 64 GB) https://geizhals.de/asus-ascent-v219542.html. Accessed 2026-11-16.
- [hardware-15] NVIDIA Developer Forums. "GB10 really does hit ~1 PFLOP NVFP4 (2:4 sparse), measured, with an open-source tool to reproduce it". https://forums.developer.nvidia.com/t/gb10-really-does-hit-1-pflop-nvfp4-2-4-sparse-measured-with-an-open-source-tool-to-reproduce-it/373618. Accessed 2026-11-16.
- [hardware-16] NVIDIA Developer Forums. "GB10 Hardware Baseline: First Direct Measurements and Findings". https://forums.developer.nvidia.com/t/gb10-hardware-baseline-first-direct-measurements-and-findings/367851. Accessed 2026-11-16.
- [hardware-17] NVIDIA Developer Forums. "The DDR bandwidth is significantly lower than the claimed 273GB/s". https://forums.developer.nvidia.com/t/the-ddr-bandwidth-is-significantly-lower-than-the-claimed-273gb-s/363238. Accessed 2026-11-16.
- [hardware-18] AMD. "AMD Ryzen AI Max+ 395" (product specifications) https://www.amd.com/en/products/processors/laptop/ryzen/ai-300-series/amd-ryzen-ai-max-plus-395.html. Accessed 2026-11-16.
- [hardware-19] AMD. "AMD Ryzen AI Halo Developer Platform with Ryzen AI Max+ PRO 495 processor" (technical specifications) https://www.amd.com/en/products/processors/desktops/ryzen/ryzen-ai-halo/ryzen-ai-max-plus-pro-495.html. Accessed 2026-11-16.
- [hardware-21] Hardware Busters, crmaris. "Framework Desktop goes to 192GB with Gorgon Halo and the price nearly doubles to $6,799". https://hwbusters.com/news/framework-desktop-goes-to-192gb-with-gorgon-halo-and-the-price-nearly-doubles-to-6799/. Accessed 2026-11-16.

- [hardware-22] geizhals.de. "GMKtec EVO-X2 (Ryzen AI Max+ 395, 128 GB, 2 TB)". https://geizhals.de/gmktec-evo-x2-a3485299.html. Accessed 2026-11-16.
- [hardware-25] NVIDIA. "Jetson Thor" (Jetson AGX Thor specifications) https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-thor/. Accessed 2026-11-16.
- [hardware-26] CNX Software, Jean-Luc Aufranc. "NVIDIA increases the price of Jetson modules and devkits by up to 101%". https://www.cnx-software.com/2026/07/22/nvidia-increases-the-price-of-jetson-modules-and-devkits-by-up-to-101/. Accessed 2026-11-16.
- [hardware-27] ServeTheHome, Ryan Smith. "NVIDIA RTX Spark N1X launch overview". https://www.servethehome.com/nvidia-rtx-spark-n1x-launch-overview-nvidia-arms-windows/. Accessed 2026-11-16.
- [hardware-28] Tech My Money, Michael John-Anyaehie. "Surface RTX Spark Dev Box pre-order ($5,999)". https://techmymoney.com/2026/10/07/surface-rtx-spark-dev-box-pre-order-5999/. Accessed 2026-11-16.
- [hardware-30] NVIDIA. "NVIDIA DGX Station" (product page, specifications) https://www.nvidia.com/en-us/products/workstations/dgx-station/. Accessed 2026-11-16.
- [hardware-31] ServeTheHome, Ryan Smith. "NVIDIA DGX Station Systems Available At Last: GB300 and GB200 Workstations For Your Desktop". https://www.servethehome.com/nvidia-dgx-station-systems-available-at-last-gb300-gb200-workstations-for-your-desktop/. Accessed 2026-11-16.
- [hardware-32] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Workstation Edition" (product page) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-6000/. Accessed 2026-11-16.
- [hardware-33] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition" datasheet (5349650, JUN26) https://www.nvidia.com/content/dam/en-zz/Solutions/products/workstations/professional-desktop-gpus/rtx-pro-6000-max-q/workstation-datasheet-blackwell-rtx-pro-6000-max-q-nvidia-us-5349650-web.pdf. Accessed 2026-11-16.
- [hardware-34] NVIDIA. "NVIDIA RTX PRO Blackwell GPU Architecture" whitepaper v1.1, Appendix A, Table 4. https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/pdf/NVIDIA-RTX-Blackwell-PRO-GPU-Architecture-v1_1.pdf. Accessed 2026-11-16.
- [hardware-35] NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Server Edition" datasheet (3850600, APR25), copy hosted by Akamai. https://www.akamai.com/site/en/documents/corporate/nvidia-rtx-pro-6000-blackwell-server-edition-datasheet.pdf. Accessed 2026-11-16.
- [hardware-36] NVIDIA. "NVIDIA RTX PRO 5000 Blackwell" datasheet (5488550), copy hosted by Open Zeka. https://openzeka.com/wp-content/uploads/2026/08/workstation-datasheet-blackwell-rtx-pro-5000-5488550-nvidia.pdf. Accessed 2026-11-16.
- [hardware-37] NVIDIA. "NVIDIA RTX PRO 5000 Blackwell" (product page) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-5000/. Accessed 2026-11-16.
- [hardware-38] NVIDIA. "NVIDIA RTX PRO 4500 Blackwell" (product page) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-4500/. Accessed 2026-11-16.
- [hardware-39] NVIDIA. "NVIDIA RTX PRO 4000 Blackwell" (product page) https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-4000/. Accessed 2026-11-16.
- [hardware-42] NVIDIA. "NVIDIA RTX Blackwell GPU Architecture" whitepaper v1.1, Appendix A, Table 3 (GeForce RTX 5090) https://images.nvidia.com/aem-dam/Solutions/geforce/blackwell/nvidia-rtx-blackwell-gpu-architecture.pdf. Accessed 2026-11-16.
- [hardware-43] NVIDIA. "NVIDIA Driver License Agreement" (GeForce), v. February 25, 2025, section 2.8. https://www.nvidia.com/en-us/drivers/geforce-license. Accessed 2026-11-16.
- [hardware-44] Notebookcheck. "Nvidia senkt Preise für GeForce RTX 5090, RTX 5080 und RTX 5070 in Europa". https://www.notebookcheck.com/Nvidia-senkt-Preise-fuer-GeForce-RTX-5090-RTX-5080-und-RTX-5070-in-Europa-so-teuer-sind-die-Grafikkarten-tatsaechlich.982114.0.html. Accessed 2026-11-16.
- [hardware-45] geizhals.de. "MSI GeForce RTX 5090" variant overview. https://geizhals.de/msi-geforce-rtx-5090-v186742.html. Accessed 2026-11-16.
- [hardware-47] geizhals.de. "NVIDIA RTX PRO 6000 Blackwell Workstation Edition, 96GB GDDR7, bulk (900-5G144-2200-000)". https://geizhals.de/nvidia-rtx-pro-6000-blackwell-workstation-edition-900-5g144-2200-000-a3474510.html. Accessed 2026-11-16.
- [hardware-48] geizhals.de. "NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (900-5G153-2500-000)". https://geizhals.de/nvidia-rtx-pro-6000-blackwell-max-q-workstation-edition-900-5g153-2500-000-a3488390.html. Accessed 2026-11-16.

- [hardware-49] geizhals.de. "NVIDIA RTX PRO 5000 Blackwell, 48GB GDDR7, retail (900-5G153-2550-000)". https://geizhals.de/nvidia-rtx-pro-5000-blackwell-900-5g153-2550-000-a3484625.html. Accessed 2026-11-16.
- [hardware-50] geizhals.de. "NVIDIA RTX PRO 5000 Blackwell, 72GB GDDR7, retail (900-5G153-2570-000)". https://geizhals.de/nvidia-rtx-pro-5000-blackwell-900-5g153-2570-000-a3687804.html. Accessed 2026-11-16.
- [hardware-51] geizhals.de. "NVIDIA RTX PRO 4500 Blackwell Workstation Edition, 32GB GDDR7, bulk (900-5G147-2250-000)". https://geizhals.de/nvidia-rtx-pro-4500-blackwell-900-5g147-2250-000-a3473415.html. Accessed 2026-11-16.
- [hardware-52] geizhals.de. "NVIDIA RTX PRO 4000 Blackwell, 24GB GDDR7, retail (900-5G147-2570-000)". https://geizhals.de/nvidia-rtx-pro-4000-blackwell-900-5g147-2570-000-a3488409.html. Accessed 2026-11-16.
- [hardware-54] AMD. "AMD Radeon AI PRO R9700" (product specifications) https://www.amd.com/en/products/graphics/workstations/radeon-ai-pro/ai-9000-series/amd-radeon-ai-pro-r9700.html. Accessed 2026-11-16.
- [hardware-55] AMD. "AMD Radeon PRO W7900" (product specifications) https://www.amd.com/en/products/graphics/workstations/radeon-pro/w7900.html. Accessed 2026-11-16.
- [hardware-57] geizhals.de. "XFX Radeon AI PRO R9700". https://geizhals.de/xfx-radeon-ai-pro-r9700-rx-97xproaiy-a3650663.html. Accessed 2026-11-16.
- [hardware-58] geizhals.de. "AMD Radeon PRO W7900". https://geizhals.de/amd-radeon-pro-w7900-a2948314.html. Accessed 2026-11-16.
- [hardware-59] Intel. "Intel Arc Pro B-Series Graphics Quick Reference Guide" (March 2026 update) https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2026-03/intel-arc-pro-b-series-graphics-quick-reference-guide-v1-0.pdf. Accessed 2026-11-16.
- [hardware-60] StorageReview, Dylan Dougherty. "Intel Arc Pro B70 Review". https://www.storagereview.com/review/intel-arc-pro-b70-review. Accessed 2026-11-16.
- [hardware-61] geizhals.de. "ASRock Arc Pro B70 Creator 32GB". https://geizhals.de/asrock-arc-pro-b70-creator-32gb-a3772447.html. Accessed 2026-11-16.
- [hardware-62] geizhals.de. "Sparkle Arc Pro B60 Passive". https://geizhals.de/sparkle-arc-pro-b60-passive-1a1-s00442300g-a3652301.html. Accessed 2026-11-16.
- [hardware-63] NVIDIA. "NVIDIA L4 Tensor Core GPU" (product page, specifications) https://www.nvidia.com/en-us/data-center/l4/. Accessed 2026-11-16.
- [hardware-64] NVIDIA. "NVIDIA L40S" (product page, specifications) https://www.nvidia.com/en-us/data-center/l40s/. Accessed 2026-11-16.
- [hardware-65] NVIDIA. "NVIDIA H100 Tensor Core GPU" (product page, specifications) https://www.nvidia.com/en-us/data-center/h100/. Accessed 2026-11-16.
- [hardware-66] NVIDIA. "NVIDIA H200" (product page, specifications) https://www.nvidia.com/en-us/data-center/h200/. Accessed 2026-11-16.
- [hardware-67] NVIDIA. "NVIDIA DGX B200" (product page, specifications) https://www.nvidia.com/en-us/data-center/dgx-b200/. Accessed 2026-11-16.
- [hardware-68] NVIDIA. "NVIDIA HGX Platform" (HGX B300, HGX B200, HGX Rubin NVL8 preliminary specifications) https://www.nvidia.com/en-us/data-center/hgx/. Accessed 2026-11-16.
- [hardware-69] NVIDIA. "NVIDIA DGX B300" (product page, specifications) https://www.nvidia.com/en-us/data-center/dgx-b300/. Accessed 2026-11-16.
- [hardware-71] GCN, Hugo Rojas. "NVIDIA Vera Rubin chips begin shipping to cloud providers as AI factory buildout accelerates". https://gcn.com/?p=20421. Accessed 2026-11-16.
- [hardware-72] AMD. "AMD Instinct MI300X" (product specifications) https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html. Accessed 2026-11-16.
- [hardware-75] AMD. "AMD Instinct MI355X" (product specifications) https://www.amd.com/en/products/accelerators/instinct/mi350/mi355x.html. Accessed 2026-11-16.
- [hardware-76] AMD. "AMD Instinct MI455X" (product specifications) https://www.amd.com/en/products/accelerators/instinct/mi400/mi455x.html. Accessed 2026-11-16.
- [hardware-77] Intel. "Intel Gaudi 3 AI Accelerator White Paper" (V1 Rev. 2), copy hosted by Presidio. https://www.presidio.com/wp-content/uploads/2025/06/gaudi-3-ai-accelerator-white-paper.pdf. Accessed 2026-11-16.

- [hardware-79] Intel press release, republished by Edge AI and Vision Alliance. "Computex: Intel Accelerates AI Everywhere, Redefines Power, Performance and Affordability". https://www.edge-ai-vision.com/2024/06/computex-intel-accelerates-ai-everywhere-redefines-power-performance-and-affordability/. Accessed 2026-11-16.
- [hardware-82] geizhals.de. "NVIDIA L4 (900-2G193-0000-001)". https://geizhals.de/nvidia-l4-900-2g193-0000-001-a3464751.html. Accessed 2026-11-16.
- [hardware-83] geizhals.de. "NVIDIA L40S (900-2G133-0080-000)". https://geizhals.de/nvidia-l40s-900-2g133-0080-000-a3127476.html. Accessed 2026-11-16.
- [hardware-84] geizhals.at. "NVIDIA RTX PRO 6000 Blackwell Server Edition (900-2G153-0000-000)". https://geizhals.at/nvidia-rtx-pro-6000-blackwell-server-edition-900-2g153-0000-000-a3492247.html. Accessed 2026-11-16.
- [hardware-86] geizhals.de. "NVIDIA H200 NVL (900-21010-0040-000)". https://geizhals.de/nvidia-h200-nvl-900-21010-0040-000-a3414253.html. Accessed 2026-11-16.
- [hardware-87] American Compute. "GPU server values" (Rack Report October 2026, data through August 2026) https://www.amcompute.com/gpu-server-values. Accessed 2026-11-16.
- [hardware-88] American Compute. "B300 vs B200" (prices as of October 2026) https://www.amcompute.com/compare/b300-vs-b200. Accessed 2026-11-16.
- [hardware-89] American Compute. "GB300 NVL72 vs HGX B300" (prices as of October 2026) https://www.amcompute.com/compare/gb300-vs-b300. Accessed 2026-11-16.
- [hardware-91] The Register, Tobias Mann. Intel Gaudi 3 launch coverage (Vision 2024) https://www.theregister.com/2024/04/09/intel_gaudi_ai_accelerator/. Accessed 2026-11-16.
- [hardware-92] CG Channel, Jim Thacker. "NVIDIA launches 72GB RTX PRO 5000 Blackwell GPU". https://www.cgchannel.com/2025/12/nvidia-launches-new-72gb-rtx-pro-5000-blackwell-gpu/. Accessed 2026-11-16.

Inference measurements (note `fonti/measurements.md`)

- [measurements-2] benches/dgx-spark/dgx-spark.md (build 7941, data captured 2026-02-05). ggml-org / llama.cpp. https://github.com/ggml-org/llama.cpp/blob/master/benches/dgx-spark/dgx-spark.md. Published: 2026-02-05 (commit 3795cc1). Accessed 2026-11-16.
- [measurements-3] benches/dgx-spark/dgx-spark.md at commit 15274c0 (build 6989, data captured 2025-11-02). ggml-org / llama.cpp. https://github.com/ggml-org/llama.cpp/blob/15274c0/benches/dgx-spark/dgx-spark.md. Published: 2025-11-10. Accessed 2026-11-16.
- [measurements-5] guide : running gpt-oss with llama.cpp (GitHub Discussion #15396), benchmark tables in the opening post. ggml-org / llama.cpp maintainers. https://github.com/ggml-org/llama.cpp/discussions/15396. Published: 2025-08-18 (tables use builds 6199 to 6310). Accessed 2026-11-16.
- [measurements-7] DGX Spark benchmark spreadsheet linked from the LMSYS review (CSV export read). LMSYS Org. https://docs.google.com/spreadsheets/d/1SF1u0J2vJ-ou-R_Ry1JZQ0iscOZL8UKHpdVFr85tNLU/edit?usp=sharing. Published: not stated (linked on 2025-10-13). Accessed 2026-11-16.
- [measurements-9] How NVIDIA DGX Spark's Performance Enables Intensive AI Tasks. NVIDIA Technical Blog. https://developer.nvidia.com/blog/how-nvidia-dgx-sparks-performance-enables-intensive-ai-tasks. Published: 2025-10-24. Accessed 2026-11-16.
- [measurements-11] NVIDIA DGX Spark performance. Ollama blog. https://ollama.com/blog/nvidia-spark-performance. Published: 2025-10-23. Accessed 2026-11-16.
- [measurements-12] Measured inference benchmarks on a single DGX Spark: same harness across Ollama, llama.cpp and vLLM (notes + data published). NVIDIA Developer Forums, community post. https://forums.developer.nvidia.com/t/measured-inference-benchmarks-on-a-single-dgx-spark-same-harness-across-ollama-llama-cpp-and-vllm-notes-data-published/379766. Published: 2026-08-10. Accessed 2026-11-16.
- [measurements-13] Qwen3.8-Flash-Next, GLM-5.3-Flash, DeepSeek-V4-Flash on 2x DGX Spark - prefill, TTFT, SWE-bench Pro. NVIDIA Developer Forums, community post. https://forums.developer.nvidia.com/t/qwen3-8-flash-next-glm-5-3-flash-deepseek-v4-flash-on-2x-dgx-spark-prefill-ttft-swe-bench-pro/382697. Published: 2026-09-08 (follow-ups 2026-09-08 and 2026-09-10). Accessed 2026-11-16.

- [measurements-20] AMD Strix Halo: Legacy Backend Benchmarks (results.json; llama.cpp builds 9187/9193). Strix Halo Toolboxes project (community). https://kyuz0.github.io/amd-strix-halo-toolboxes/. Published: 2026-05-18 (data captured 2026-05-16). Accessed 2026-11-16; rechecked 2026-11-17 (results.json: five backends).
- [measurements-21] Local LLM Benchmarks: Speed results (data/results.json). Local LLM Benchmarks, Strix Halo Toolboxes project (community). https://local-llm-benchmarks.dev/speed.html. Published: continuous; runs used here dated 2026-09-01 to 2026-10-26. Accessed 2026-11-16.
- [measurements-24] AA-AgentPerf-Local: Benchmarking local AI agents on laptops and workstations. Artificial Analysis. https://artificialanalysis.ai/articles/aa-agentperf-local. Published: 2026-09-29. Accessed 2026-11-16; rechecked 2026-11-17.
- [measurements-25] Laptop & Workstation Inference Systems (results page, prices as of October 2026). Artificial Analysis. https://artificialanalysis.ai/hardware-inference-stack/laptops-workstations. Published: not stated (page read 2026-11-16). Accessed 2026-11-16; rechecked 2026-11-17 (AMD system listed as Ryzen AI Max+ 395, 128 GB; runs labelled "(Reasoning)" and "(Xhigh)").
- [measurements-26] MLPerf Inference v6.1 results (summary.csv and per-submission result_summary.json). MLCommons. https://github.com/mlcommons/inference_results_v6.1. Published: 2026-09-16. Accessed 2026-11-16; rechecked 2026-11-17 (summary.csv; result_summary.json of the two Atlas Inference Edge Agentic runs).
- [measurements-28] MLPerf Inference: gpt-oss-120b benchmark description. MLCommons. https://mlcommons.org/2026/03/mlperf-inference-gpt-oss/. Published: 2026-03-24. Accessed 2026-11-16.
- [measurements-29] MLPerf Inference v6.1: 5.7x Per-Accelerator Gains, a 512-GPU Run, and Vera Rubin's First Peer-Reviewed Numbers. StorageReview. https://www.storagereview.com/news/mlperf-inference-v6-1-5-7x-per-accelerator-gains-a-512-gpu-run-and-vera-rubins-first-peer-reviewed-numbers. Published: 2026-09-16. Accessed 2026-11-16; rechecked 2026-11-17 (describes the Edge Agentic workload; gives 20.1 and 19.63 tok/s, no per-turn values).
- [measurements-30] InferenceMAX: Open Source Inference Benchmarking. SemiAnalysis. https://newsletter.semianalysis.com/p/inferencemax-open-source-inference. Published: 2025-10-09. Accessed 2026-11-16.
- [measurements-32] Running gpt-oss-120b on H100 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-h100. Published: live page; runs 2026-03-27 to 2026-05-17. Accessed 2026-11-16.
- [measurements-33] Running gpt-oss-120b on H200 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-h200. Published: live page; runs 2025-12-15 to 2026-05-30. Accessed 2026-11-16.
- [measurements-34] Running gpt-oss-120b on B200 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-b200. Published: live page; runs 2026-02-05 to 2026-05-31. Accessed 2026-11-16.
- [measurements-35] Running gpt-oss-120b on MI300X (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-mi300x. Published: live page; runs 2026-03-09 to 2026-04-21. Accessed 2026-11-16.
- [measurements-36] Running gpt-oss-120b on MI355X (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/gptoss-120b-on-mi355x. Published: live page; runs 2026-01-14 to 2026-06-05. Accessed 2026-11-16.
- [measurements-40] Running Qwen3.5 397B on H100 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-h100. Published: live page; runs 2026-07-05 to 2026-10-20. Accessed 2026-11-16.
- [measurements-41] Running Qwen3.5 397B on H200 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-h200. Published: live page; runs 2026-03-04 to 2026-10-28. Accessed 2026-11-16.
- [measurements-42] Running Qwen3.5 397B on B200 (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-b200. Published: live page; runs 2026-02-20 to 2026-11-08. Accessed 2026-11-16.
- [measurements-43] Running Qwen3.5 397B on MI355X (InferenceX run page). SemiAnalysis InferenceX. https://inferencex.semianalysis.com/run/qwen-3-5-on-mi355x. Published: live page; runs 2026-02-17 to 2026-11-09. Accessed 2026-11-16.
- [measurements-51] SGLang vs DwarfStar vs vLLM: DeepSeek 4 on RTX Pro 6000. HelixML. https://helix.ml/blog/running-ds4-on-rtx-pro-6000. Published: 2026-08-14. Accessed 2026-11-16.

Benchmarks (note `fonti/benchmarks.md`)

- [benchmarks-1] Artificial Analysis Intelligence Index v5 (launch article). Artificial Analysis. https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5. Accessed 2026-11-16.
- [benchmarks-3] Open weights models comparison (Intelligence Index v5.0.1, parameters). Artificial Analysis. https://artificialanalysis.ai/models/open-source. Accessed 2026-11-16.
- [benchmarks-4] Comparison of AI Models; Models leaderboard (Intelligence Index v5.0.1). Artificial Analysis. https://artificialanalysis.ai/models ; https://artificialanalysis.ai/leaderboards/models. Accessed 2026-11-16.
- [benchmarks-5] Model pages: gpt-oss-120b, gpt-oss-20b, Gemma 4 31B, Qwen3.6 35B A3B. Artificial Analysis. https://artificialanalysis.ai/models/gpt-oss-120b ; https://artificialanalysis.ai/models/gpt-oss-20b ; https://artificialanalysis.ai/models/gemma-4-31b ; https://artificialanalysis.ai/models/qwen3-6-35b-a3b. Accessed 2026-11-16.
- [benchmarks-10] NVIDIA launches Nemotron 3.5 Lightning. Artificial Analysis. https://artificialanalysis.ai/articles/nemotron-3-5-lightning-launch. Accessed 2026-11-16.
- [benchmarks-11] Changelog (entries 30 Oct to 16 Nov 2026). Artificial Analysis. https://artificialanalysis.ai/changelog. Accessed 2026-11-16.
- [benchmarks-13] Multilingual Index: Italian; Multilingual Index overview (Global-MMLU-Lite based). Artificial Analysis. https://artificialanalysis.ai/models/multilingual/italian ; https://artificialanalysis.ai/models/multilingual. Accessed 2026-11-16.
- [benchmarks-14] Capabilities & benchmarking hub, LLM benchmark data (benchmark_data.zip, CC BY 4.0); files used: swe_bench_verified.csv (run by Epoch), terminalbench_external.csv, os_world_external.csv, osworld_2_external.csv, model_metadata.csv. Epoch AI. https://epoch.ai/benchmarks ; https://epoch.ai/data/benchmark_data.zip. Accessed 2026-11-16.
- [benchmarks-15] Epoch Capabilities Index (ECI), eci_scores.csv in the benchmark data download. Epoch AI. https://epoch.ai/eci ; https://epoch.ai/data/benchmark_data.zip. Accessed 2026-11-16; rechecked 2026-11-17 (eci_scores.csv dated 2026-11-16).
- [benchmarks-17] Terminal-Bench 4.0 (mini-swe-agent, avg@3). Vals AI. https://www.vals.ai/benchmarks/terminal-bench-4. Accessed 2026-11-16.
- [benchmarks-18] SWE-bench Verified (archived: no longer run on new releases). Vals AI. https://www.vals.ai/benchmarks/swebench. Accessed 2026-11-16.
- [benchmarks-23] SWE-bench leaderboards and data file leaderboards.json (bash-only board removed from the site on 2026-09-01). SWE-bench maintainers. https://www.swebench.com/ ; https://github.com/SWE-bench/swe-bench.github.io. Accessed 2026-11-16.
- [benchmarks-24] Why we no longer evaluate SWE-bench Verified (automated access returned HTTP 403; content as summarised in benchmarks-25). OpenAI. https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/. Accessed 2026-11-16.
- [benchmarks-25] Summary of the SWE-bench Verified retirement (secondary source). Pebblous blog. https://blog.pebblous.ai/blog/swe-bench-verified-retired/en/. Accessed 2026-11-16.
- [benchmarks-32] Berkeley Function Calling Leaderboard (BFCL) V4 and data_overall.csv. UC Berkeley (Gorilla project) https://gorilla.cs.berkeley.edu/leaderboard.html. Accessed 2026-11-16.
- [benchmarks-34] OSWorld-Verified official results (osworld_verified_results.xlsx, evaluated by the OSWorld team under unified settings). XLANG Lab, University of Hong Kong. https://osworld-v1.xlang.ai/. Accessed 2026-11-16.
- [benchmarks-42] Qwen3.8-27B model card. Qwen team (Alibaba) https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 2026-11-16.
- [benchmarks-45] Qwen3.6-35B-A3B model card. Qwen team (Alibaba) https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed 2026-11-16.
- [benchmarks-48] Muse Glimmer-30B model card. Meta. https://huggingface.co/meta-models/Muse-Glimmer-30B. Accessed 2026-11-16.
- [benchmarks-49] gpt-oss-120b & gpt-oss-20b Model Card (arXiv 2508.10925). OpenAI. https://arxiv.org/abs/2508.10925. Accessed 2026-11-16.
- [benchmarks-54] Nemotron 3.5 Super and Lightning technical report (Table 10, Sections 3.8 and 1) and project page. NVIDIA. https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3.5-Super-and-Lightning-Technical-Report.pdf ; https://research.nvidia.com/labs/nemotron/Nemotron-3.5-Super-and-Lightning/. Accessed 2026-11-16.

- [benchmarks-55] Holo3-35B-A3B model card. H Company. https://huggingface.co/Hcompany/Holo3-35B-A3B. Accessed 2026-11-16.
- [benchmarks-56] Hugging Face Hub API model metadata (licence tag, safetensors parameter count, repository creation date) for the repositories in section 3. Hugging Face. https://huggingface.co/api/models/{repo_id}. Accessed 2026-11-16.
- [benchmarks-57] Benchmarking EngGPT2-16B-A3B against Comparable Italian and International Open-source LLMs (arXiv 2605.07731v2), Tables 16 and 17. ENGINEERING Ingegneria Informatica. https://arxiv.org/abs/2605.07731. Accessed 2026-11-16.

Method references added by the author

- [method-1] llama.cpp, tools/quantize/README.md: bits per weight of each quantization type measured on Llama-3.1-8B (Q4_K_M 4.8944, Q8_0 8.5008, IQ4_XS 4.4597, F16 16.0005). ggml-org. https://github.com/ggml-org/llama.cpp/blob/master/tools/quantize/README.md. Accessed 2026-11-16.
- [method-2] E. Alvarez et al., "Introducing NVFP4 for Efficient and Accurate Low-Precision Inference", NVIDIA Technical Blog, 2025-06-24 (NVFP4: one FP8 E4M3 scale per 16 values, 4.5 bits per value; MXFP4: one power-of-two scale per 32 values). Vendor publication. https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/. Accessed 2026-11-16.
- [method-3] vLLM documentation, "vllm serve" CLI reference: --gpu-memory-utilization defaults to 1.0 with the extensible KV cache and 0.92 otherwise. https://docs.vllm.ai/en/latest/cli/serve.html. Accessed 2026-11-16.
- [method-4] A. Mekala, A. Atmakuru, Y. Song, M. Karpinska, M. Iyyer, "Does quantization affect models' performance on long-context tasks?", arXiv:2505.20276 (v1 2025-05-26, v3 2025-09-20). Independent academic study. https://arxiv.org/abs/2505.20276. Accessed 2026-11-16.
- [method-5] P. Migdał, "Benchmarking Qwen3.8 27B quantizations: 4-bit holds up, 1-bit collapses", Quesma blog, 2026-08-26 (llama.cpp, BF16 vs Q8_0 vs Q4_K_M and lower, GPQA Diamond, IFBench, Terminal-Bench 2.1). Independent, single model, blog. https://quesma.com/blog/qwen38-27b-quantizations-benchmarked/. Accessed 2026-11-16.
- [method-6] SparkTech, "SparkTest on open-weight models: measurements on a CPU test server", 17 November 2026, companion document of this study (section 5.7). Measured by SparkTech; raw data and scripts available on request. https://autotest.sparktech.it/docs/sparktest-open-weight-measurements-en.pdf.

Verification

On 16 and 17 November 2026 two independent reviews checked this study; the corrections were applied on 17 November 2026.

- **Figures and sources.** 91 claims, covering more than 250 individual values, were checked against the cited primary sources and, where possible, a second source; all 218 distinct URLs of the references were opened (none missing; 6 refused automated clients and 6 hit a rate limit, and these were checked by other means). Result: 82 confirmed, 7 wrong, 1 not verifiable in the cited source, 1 cited a source that does not contain the values. All 9 were corrected: three Epoch Capabilities Index values that were reported as not found (table 3.2), the SWE-bench Verified range of section 8.2, the label of GLM-4.7-Flash's active parameters, the FP4 figure of the 48 GB RTX PRO 5000, the Ryzen AI Max+ 395 prefill range at 32k depth (234 to 653 tok/s, five backends instead of three), the source of the MLPerf Edge Agentic values (table 5.6) and the source of the GLM-5.3 parameter counts. None of these changes a conclusion.
- **Consistency, labels and form.** The calculator was rerun (identical output), tables 5.1, 5.2, 5.3, 6 and 9.3 were compared with it cell by cell (no difference), five rows of table 6 were recomputed with an independent implementation of the formulas, and the 228 citations were matched to the references. 34 points were corrected: 17 consistency issues (among them the 3 internal inconsistencies also found by the first review), 2 missing formula descriptions, 6 labelling issues, 8 unit or form issues and 1 sentence about the SparkNode. Clarity additions: the box for IT teams, the glossary, the choice by size of section 6.4 and short notes in sections 1 to 8.

- **What changed in the recommendations.** Every size verdict now also checks codify prompts with application documentation at the cap, and table 6 gives the size each configuration is usable up to. Class 1: Ryzen AI Max+ 395 is usable up to size M, not L; on a compact system 117B to 125B models reach size M on GB10 and size S on Ryzen AI Max+ 395 (gpt-oss-120b only), within the hard limits; with documentation at the cap, GB10 and the 32 GB cards meet the targets up to size S. Table 12 of `out/tabelle.md` shows how the class 1 sizes move with the decode efficiency.

In total, 43 points were corrected out of the 91 claims and the 5 tables, 228 citations and 218 URLs checked.

Updates of 17 November 2026, after the verification.

- **No new codification after a failed run; repairs keep values.** The agent no longer codifies a test again after a failed run, and its scenario check now also rejects a repair that changes an expected value or a typed value [workload-serving-67]. Section 1 was aligned; the night scenarios were recomputed (the release night with 1,500 tests went from 833 to 1,635 calls to 593 to 1,155) and the calculator was rerun: only the night capacity changed (last column of table 6, sections 6.1, 6.2 and 9.4), no size verdict.
- **The request of a customer endpoint.** The agent's request now names its JSON schema and carries `chat_template_kwargs` and `reasoning_effort`: the summary and sections 7.1 to 7.4 were aligned.
- **SparkTech's own measurement.** Section 5.7 and point 7 of the summary add the first measurement of SparkTest's tasks on open-weight models, checked by an independent review before publication [method-6].