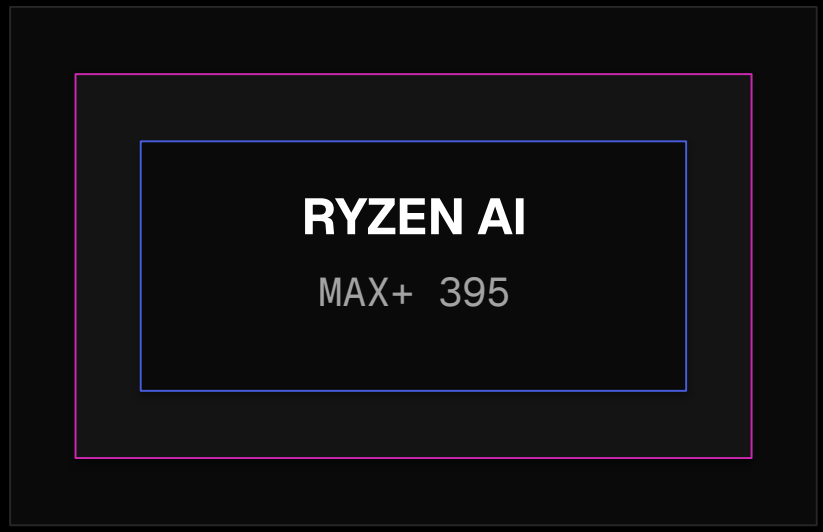
METRUM AI

Which Local LLM Stack Fits a Strix Halo Team?

AMD Ryzen AI Max+ 395
128 GB unified memory



NOVEMBER 2026

Evaluation system courtesy of Bryant Kreger @ Dell

Powering Enterprise AI From Insight to Infrastructure.

© 2026 Metrum AI, Inc.

THE ANSWER

What should a team run?

llama.cpp on ROCm is the best measured starting point.

FASTEST

153

GPT-OSS-20B · total tok/s at 16 users

BALANCED

131

Qwen3.6-35B-A3B Q4 · total tok/s at 16 users

QUALITY

42/50

gemma-4-31B · chat score

Start with ROCm for concurrent users; select the model against the quality and latency needs of your workload.

© 2026 Metrum AI, Inc.

MODEL SELECTION

Which model fits the job?

Speed, balance, chat quality and model size lead to different picks.

FASTEST	**GPT-OSS-20B**	153 tok/s total · 16 users
BALANCED	**Qwen3.6-35B-A3B Q4**	131 tok/s total · 16 users
BEST CHAT	**gemma-4-31B**	42/50 chat · 66.7 tok/s total at 16 users
LARGEST THAT FITS	**Qwen3.5-122B-A10B**	19.1 tok/s at 1 user · 39/50 chat

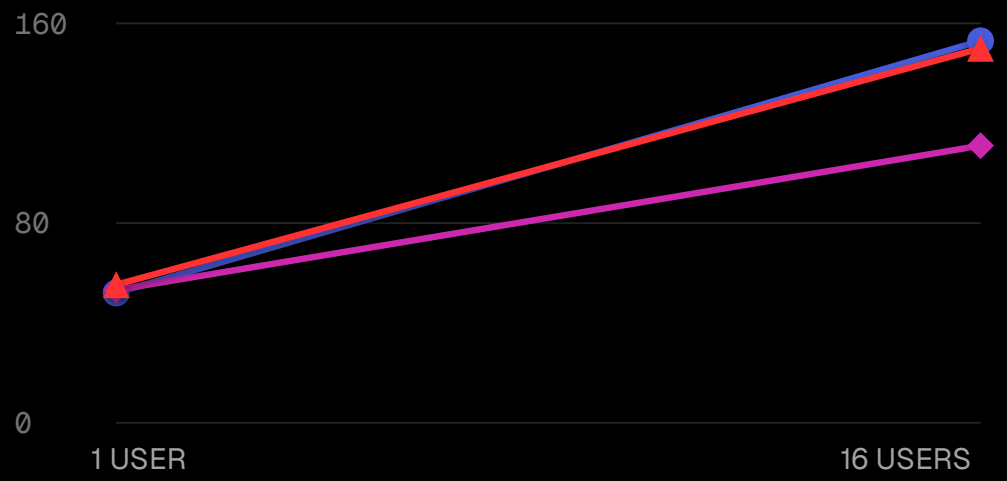
© 2026 Metrum AI, Inc.

FRAMEWORK SCALING

What changes when a team arrives?

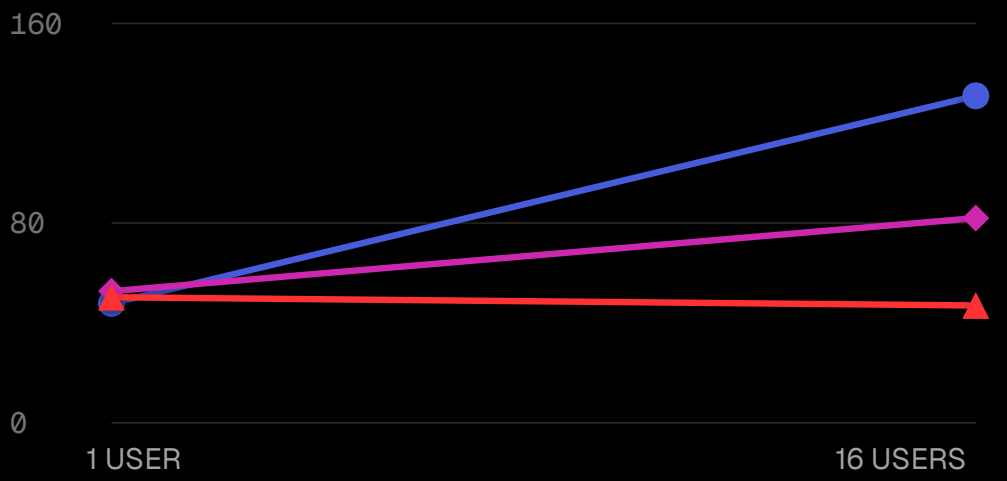
Framework throughput diverges as concurrency rises.

GPT-OSS-20B · MXFP4



ROCm	52.0 → 153
Vulkan	53.2 → 111
Ollama	55.3 → 150

Qwen3.6-35B-A3B · Q4



ROCm	47.8 → 131
Vulkan	52.7 → 82
Ollama	50.3 → 47

Output tok/s total. Ollama keeps prefix caching; llama.cpp controls do not.

© 2026 Metrum AI, Inc.

GB10 COMPARISON

Is GB10 faster?

With the same model file and runtime, the answer depends on load.

STRIX HALO · 1 USER

53

tok/s · Qwen3.6-35B-A3B Q4

GB10 · 1 USER

70

tok/s · same model and runtime

16 USERS · CHAT/CODE

~150

Halo: 150 · GB10: 151–175 total tok/s

Quality was equal within noise. The ~3× gap cited elsewhere used vLLM and GB10 NVFP4.

From our like-for-like GB10 comparison: same model file, same llama.cpp runtime.

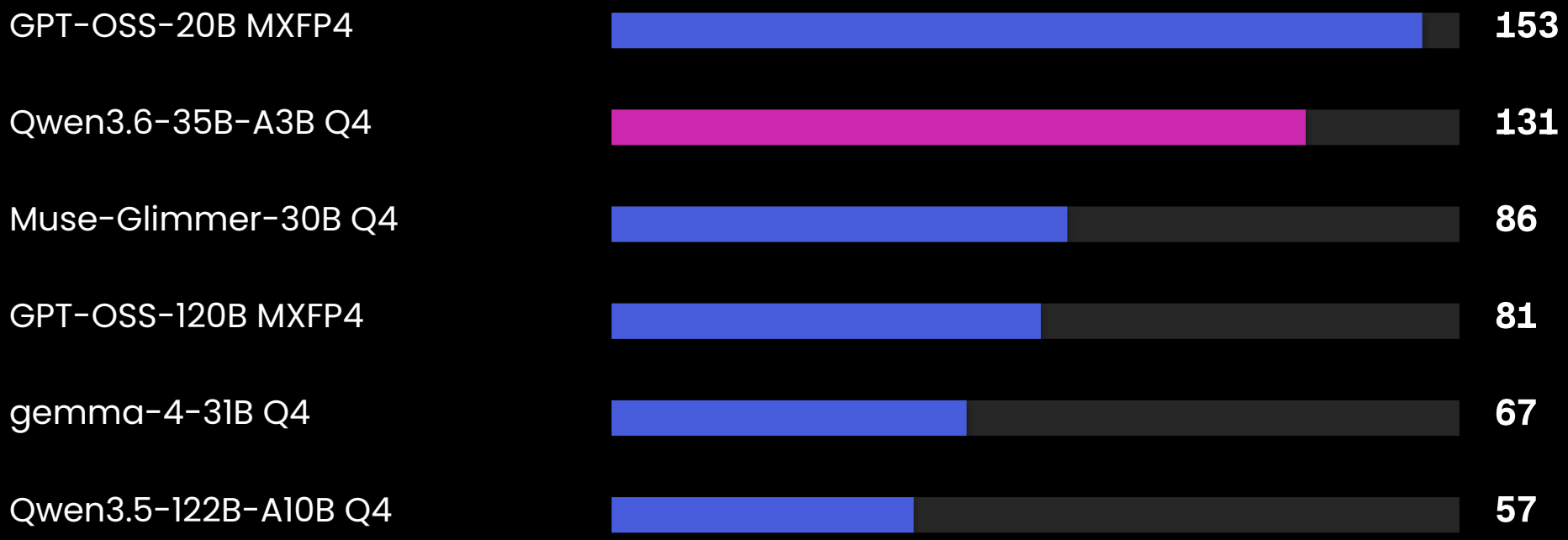
© 2026 Metrum AI, Inc.

THROUGHPUT

Where does throughput go?

Model choice changes the aggregate rate at 16 users.

`llama.cpp ROCm | output tok/s, all users`

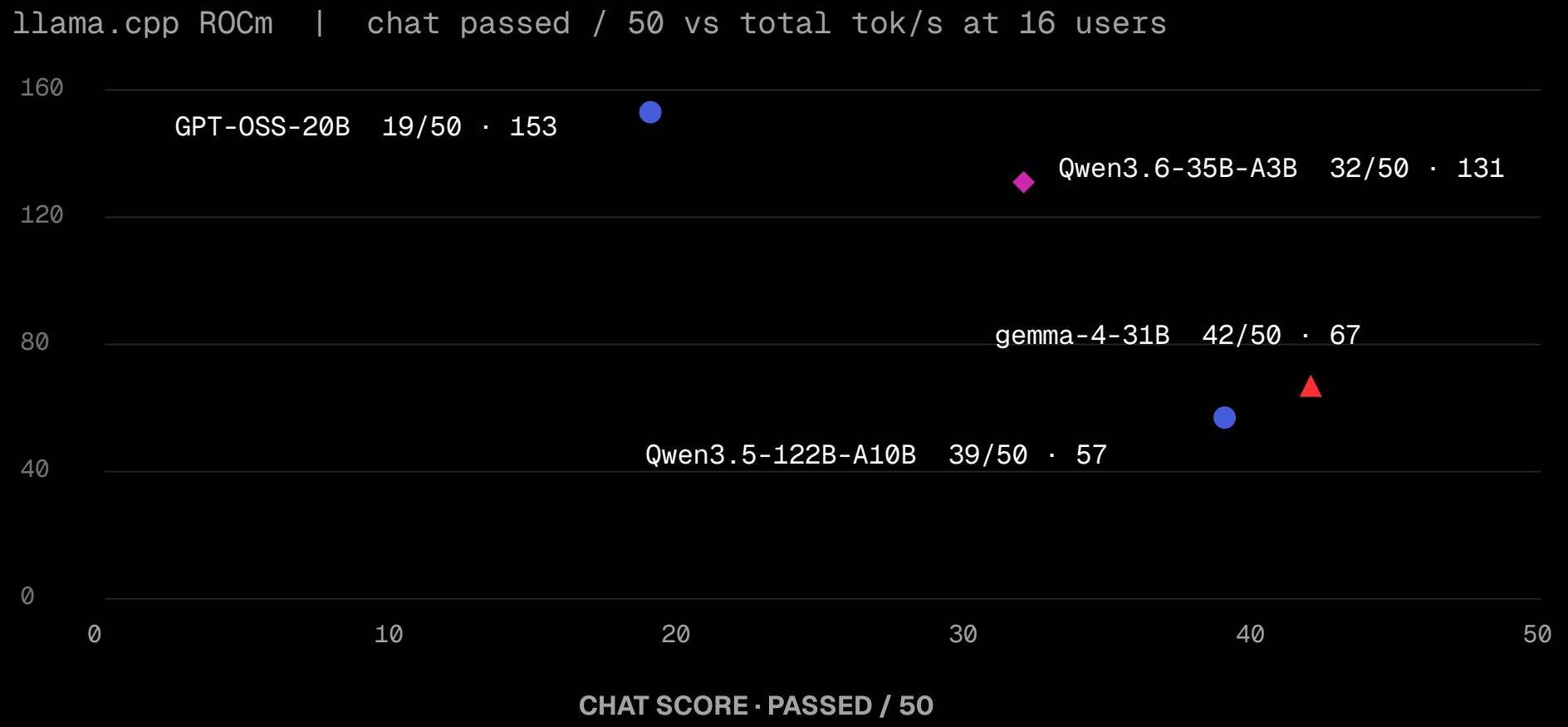


© 2026 Metrum AI, Inc.

QUALITY

Can fast also be good?

Chat scores and throughput reveal different trade-offs.



Qwen3.6 quality was measured on an earlier llama.cpp version than its speed.

© 2026 Metrum AI, Inc.

EFFICIENCY

How much work per joule?

Measure throughput against whole-socket power.

`llama.cpp ROCm | tok/J and peak memory`

GPT-OSS-20B	4.74	19 GB	32 users
Qwen3.6-35B-A3B	3.32	29 GB	32 users
gemma-4-31B	1.48	50 GB	16 users
Qwen3.5-122B-A10B	1.24	83 GB	16 users

Socket power is whole SoC power, not discrete GPU power. Peak memory is reported in GB.

© 2026 Metrum AI, Inc.

SERVING BEHAVIOR

Why do Qwen requests queue?

In this Ollama version, the tested Qwen families used one slot.

QWEN3.6 Q4

1 slot

Ollama v0.40.13

One request at a time.

The Qwen 3.5, 3.6 and 3.8 families used a single slot in this test. Plan for queues when several users submit requests together.

For GPT-OSS-20B, Ollama reached 150 tok/s total at 16 users.

© 2026 Metrum AI, Inc.

HARDWARE

What hardware was tested?

A Ryzen AI developer platform with unified memory.

AMD Ryzen AI Max+ 395

128 GB unified memory

Local inference on one AMD Ryzen AI developer platform. The graphic is abstract, not a product image.



© 2026 Metrum AI, Inc.

Why can a larger model be faster?

A mixture of experts activates only selected experts for each token.

QWEN3.5-122B-A10B

19.1

tok/s · 1 user · llama.cpp ROCm

QWEN3.8-27B DENSE

11.3

tok/s · 1 user · llama.cpp ROCm

The router selects experts per token; shared weights still participate. Other factors also affect speed.

Source: huggingface.co/docs/transformers/v5.11.0/en/experts_interface

© 2026 Metrum AI, Inc.

Why does batching help a team?

A model step can serve multiple active requests at once.

ONE SLOT	One active request; later requests wait in a queue.
CONTINUOUS BATCH	Active requests share model steps; finished slots can be refilled.
TRADE-OFF	More aggregate throughput does not guarantee faster individual responses.

Sources: huggingface.co/docs/transformers/main/continuous_batching_architecture

developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/

© 2026 Metrum AI, Inc.

METHOD

How were the results measured?

One client generated load; separate tools scored quality and telemetry.

SPEED	Bench-CLI: streaming request records; 1, 4 and 16 users, plus 32 where admitted.
QUALITY	promptfoo: frozen hard-tier tasks, automatic answer checks and executed code tests.
TELEMETRY	all-smi: SoC socket power and peak memory; tok/J uses throughput and mean power.
CAVEAT	Ollama prefix caching remains enabled; llama.cpp prompt caching is disabled.

© 2026 Metrum AI, Inc.

MEASURED WITH METRUM AI BENCH-CLI

What did the measurement run?

Metrum AI Bench-CLI issued the same command shape.

```
metrum-ai-bench-cli-strategic \
  --url http://localhost:8080/v1/chat/completions --api-key dummy \
  --model qwen3.5-122b-q4xl --prompts prompts.jsonl --max-tokens 384 \
  --extra-body-json '{"temperature":0.7,"top_p":0.8,"top_k":20,"min_p":0.0,
      "repeat_penalty":1.0,"presence_penalty":1.5,"seed":7,"cache_prompt":false,
      "chat_template_kwargs":{"enable_thinking":false}}' \
  --sweep 16 --sweep-by concurrency --requests-per-stage 32 --warmup-requests 16 \
  --timeout-seconds 3600 --seed 7 --streaming \
  --csv requests.csv --ndjson events.ndjson
```

Streaming: token timing · Sweep: concurrency · Warmup: excluded

github.com/metrum-ai/bench-cli

© 2026 Metrum AI, Inc.

What do the options control?

Streaming, concurrency and records make the results interpretable.

`--streaming`	Separates first-token and decode timing.
`--sweep / --sweep-by`	Controls concurrency levels.
`--requests-per-stage`	Sets the sample count at each level.
`--warmup-requests`	Excludes model and cache warmup.
`--prompts / --seed`	Selects a fixed seeded prompt set.
`--extra-body-json`	Sets sampling and cache behavior.
`--max-tokens 1`	Isolates prefill with long prompts.
`--csv / --ndjson`	Records requests and events.

© 2026 Metrum AI, Inc.

COVERAGE

Where are the gaps?

Missing runs and timeouts are not zero scores.

MODEL FIT	Muse Spark 816B: smallest stated GGUF exceeds 128 GB memory.
HIGH CONCURRENCY	Some 32-user tests were not admitted after the 16-user gate.
QUALITY	Some quality suites timed out; partial cases were not scored.
REFERENCES	Some older vLLM reference rows lack comparable telemetry or quality.

© 2026 Metrum AI, Inc.

How can deployment scale?

Start desk-side; plan a separate data-center path as needed.

DESK-SIDE

Developer platform

Validate the model locally, then test concurrency and queues.

DATA CENTER

Dell PowerEdge XE9680

Dell describes this server with AMD Instinct MI300X accelerators.

Also from Metrum AI: Router, open source

`github.com/metrum-ai/router`

Source: dell.com/en-us/infohub/articles/dell-poweredge-xe9680-servers-with-amd-instinct-mi300x-accelerators-genai-with-llama-3-1-405b-llms

© 2026 Metrum AI, Inc.

Where can you go deeper?

Full report and evidence are available from Metrum AI.

Metrum AI Bench-CLI · open source

`github.com/metrum-ai/bench-cli`

Quality: promptfoo · Telemetry: all-smi

Also from Metrum AI: Router · github.com/metrum-ai/router

Evaluation system courtesy of Bryant Kreger @ Dell

Powering Enterprise AI From Insight to Infrastructure.

AMD, Ryzen, Radeon and Instinct are trademarks of Advanced Micro Devices, Inc. Dell and PowerEdge are trademarks of Dell Inc. NVIDIA and GB10 are trademarks of NVIDIA Corporation. Other names are trademarks of their owners.

© 2026 Metrum AI, Inc.