METRUM AI

Current 9–35B Models Run on the Ryzen AI NPU

The 35B MoE is the one worth using.

Speed, quality and proof of NPU work | Oct 28, 2026

Dell Pro Precision 5 14S	Ryzen AI 9 HX PRO 475 · XDNA 2 NPU
Radeon 890M iGPU · 64 GB	Windows 11 Pro 26200 · FastFlowLM 1.0.10

© 2026 Metrum AI, Inc.

Good NPU quality. The iGPU is faster for the same model.

58/60	13.5 tok/s	74–98%	2.74 s
Qwen3.6-35B-A3B quality Cloud references: 55/60	Qwen3.6-35B-A3B NPU decode iGPU: 19.8 tok/s	NPU busy share across NPU model runs; iGPU at 0%	Whisper per clip on NPU CPU: 4.33 s

Qwen3.6-35B-A3B generated 13.5 tokens/s with the NPU busy 74% of the time and the iGPU idle. The same model as a GGUF build on the iGPU reached 19.8 tokens/s, with first token in 5.5 s instead of 16.2 s, at nearly the same quality (57/60).

Dense Gemma 4 12B and Qwen3.8-27B are slow; the 27B was too slow to benchmark. gpt-oss-20B (MoE) requires reasoning: 9 of 10 benchmark answers ran out of tokens. Whisper: 2.74 s per clip on NPU versus 4.33 s on CPU, with exact transcripts on three short clips.

© 2026 Metrum AI, Inc.

Use the NPU selectively in edge-first routing.

No showstopper

Every failure had an identified cause and was worked around or recorded as a limit.

Best local model: Qwen3.6-35B-A3B

Matched or beat the cloud references on this suite. Use the NPU when the iGPU is needed elsewhere; power was not measured.

Chat waits on the first token

NPU: 4.7–16 s for a ~1,000-token prompt, versus 0.64 s for gpt-4o-mini. Better suited to background and batch work.

Other models: distinct limits

Dense: Gemma 4 12B, 5 tok/s; Qwen3.8-27B, under 1 tok/s. gpt-oss-20B (MoE): mandatory reasoning; 9 of 10 benchmark answers ran out of tokens.

Supervise FastFlowLM

Stops accepting connections after 10 are left open. Restarting fixed every case.

Speech-to-text: ready for a pilot

Faster than CPU with exact transcripts on these clips. Broader accuracy testing is still needed.

© 2026 Metrum AI, Inc.

The platform works; NPU serving has known limits.

Worked → failed / cause → workaround → practical bottom line

Ticket	Worked	Did not work / cause	Workaround and bottom line
P0 · Platform	Driver, firmware and FastFlowLM; xrt-smi from driver folder.	Stale System32 xrt-smi crashes. AMD package is older. SDK 1.8 downloads reset.	Use working tool; do not downgrade. WORKS. SDK 1.8 remains a gap.
P1 · NPU LLMs	All five families loaded; device busy time and owner measured.	Early stops; 27B under 1 tok/s; gpt-oss reasoning exhausts cap; connection limit.	Use actual returned tokens; bounded subsets; separate thinking row; restart and rerun. WORKS WITH LIMITS.
P1 · iGPU LLMs	Qwen and Gemma: 10/10; exact 256-token outputs.	Gemma fenced 7/8 JSON answers; strict checks fail.	Keep strict score. Production: response_format or strip fences. WORKS.

© 2026 Metrum AI, Inc.

Extended evals and speech recovered; scope cuts are not failures.

Ticket	Worked / limit	Workaround and bottom line
P1 · Extended evals	46 more cases per top model and backend; retrieval 6/6 on all four runs. First NPU attempt hit connection limit.	Restarted and reran. All four runs fully graded. WORKS.
P3 · Whisper	NPU and CPU 10/10, zero word errors. CPU first warm-up: HTTP 500 from faster-whisper / PyAV mismatch.	Read WAV directly, no resampling. CPU rerun solved it; NPU unchanged. WORKS; only three clean clips.
P2, P4, P5, P6	Embeddings, image generation, vision and vision-language: no measurements.	NOT RUN — operator cut for time.
P7 · Power	Device power counters read 0; package counter units not validated.	NOT RUN — no watt or tokens-per-joule claims.
P8 · Router	Skipped by operator.	NOT RUN — Phase 1 results still apply.

© 2026 Metrum AI, Inc.

Leaderboard: every model, including diagnostic rows.

TTFT = time to first token. Decode and prefill are tokens/s; prefill columns are ~1k / ~4k prompts.

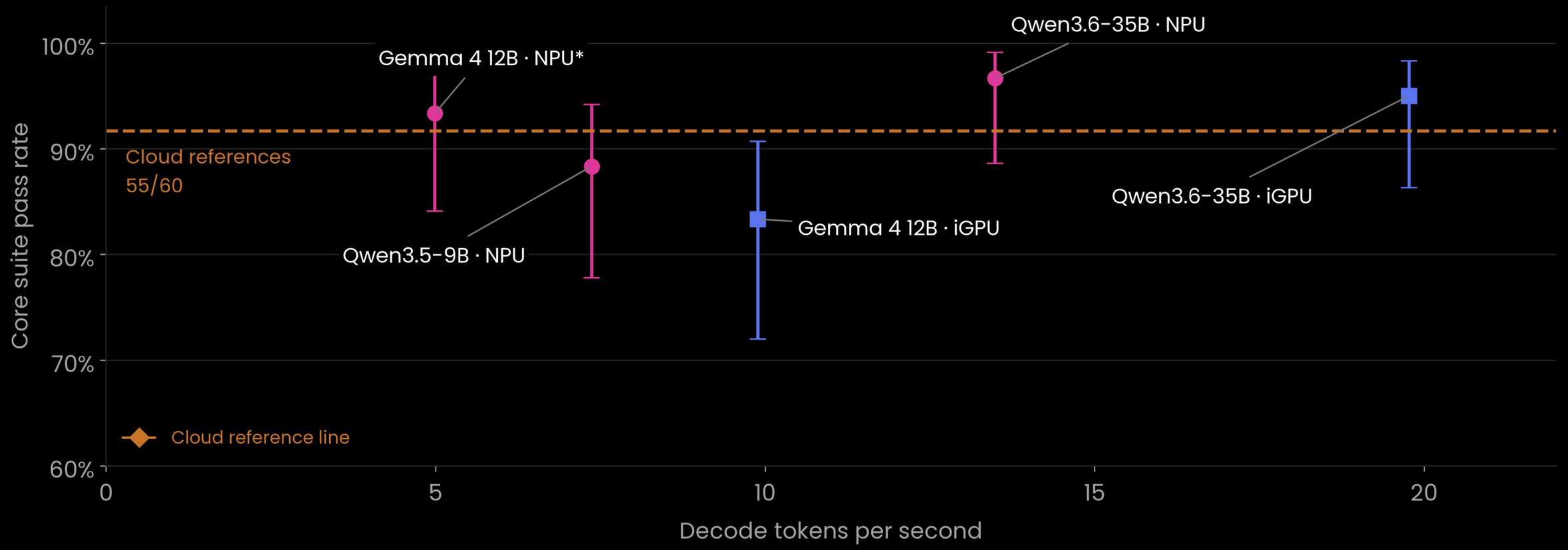
Model	Backend / quant	OK	TTFT p50	Decode	Prefill 1k / 4k	Quality · 95% CI	Device proof	Qualification
Qwen3.6-35B-A3B (MoE)	NPU · Q4_K	10/10	16.19 s	13.49	64 / 129	58/60 (88.6-99.1%)	NPU 73.7% iGPU 0%	Stops early: mean 204 of 256 tokens
Qwen3.6-35B-A3B (MoE)	iGPU · UD-Q4_K_XL	10/10	5.52 s	19.77	189 / 255	57/60 (86.3-98.3%)	NPU 0% iGPU 100%	Exact 256 tokens: PASS
Gemma 4 12B	NPU · Q4_0	9/10	6.16 s	4.99	174 / 242	56/60 (84.1-97.4%)	NPU 96.7% iGPU 0%	Diagnostic: 1 connection error
Gemma 4 12B	iGPU · Q4_K_M	10/10	12.94 s	9.89	80 / 101	50/60 (72.0-90.7%)	NPU 0% iGPU 100%	7 JSON answers wrapped in code fences
Qwen3.5-9B	NPU · Q4_K	10/10	4.67 s	7.37	223 / 313	53/60 (77.8-94.2%)	NPU 95.2% iGPU 0%	Stops early: mean 160 of 256 tokens
gpt-oss-20B (MoE)	NPU · Q4_1 reasoning on	1/10	102.5 s	n/a	178 / n/a	24/26 (75.9-97.9%)	NPU 97.0% iGPU 0%	9 answers ran out of tokens while reasoning
Qwen3.8-27B dense, MTP	NPU	2 done	n/a	< 1	n/a	27/28 graded, 30-case subset	NPU busy 520 s of 547 s	216 s and 257 s per answer. Stopped at time limit
gpt-4o-mini	Cloud reference	10/10	0.64 s	92.88	1,699 / n/a	55/60 (81.9-96.4%)	No local work	Remote hardware unknown
gpt-6-luna	Cloud reference	0/10	n/a	n/a	n/a	55/60 (81.9-96.4%)	No local work	HTTP 400 from bench-cli: #357

Fewer than 10 successful requests = diagnostic. NPU exact-length checks fail; speed uses returned tokens. Quantization/export differ across backends. Low-N percentiles are descriptive.

© 2026 Metrum AI, Inc.

Only Qwen3.6-35B Is Both Fast and Accurate on the NPU

Circles: NPU | Squares: iGPU | Diamond + dashed line: cloud references

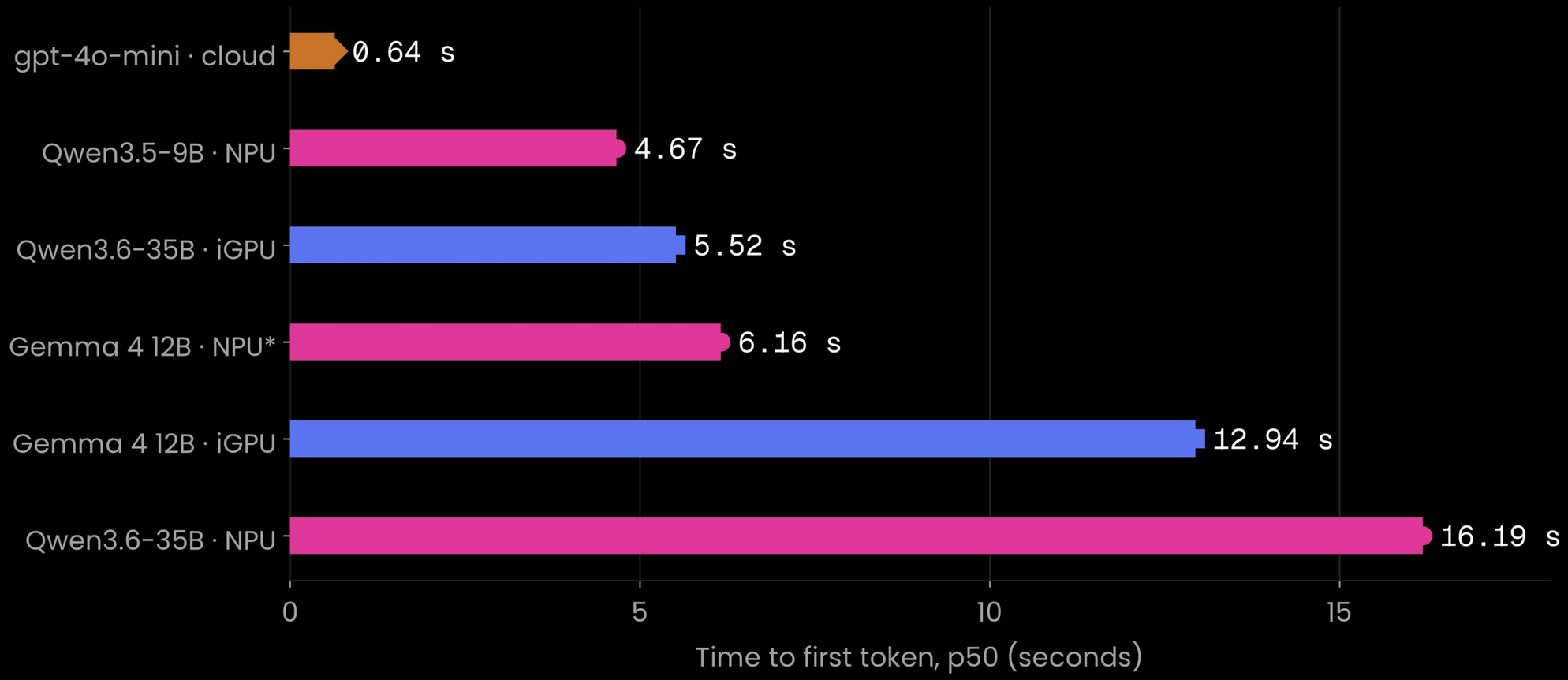


* Gemma NPU speed is diagnostic (9/10). gpt-oss and Qwen3.8 omitted: no valid decode rate. gpt-4o-mini: 92.88 tok/s, off this scale.

© 2026 Metrum AI, Inc.

First-token latency limits interactive NPU chat.

Same ~1,000-token prompt setting. Circles: NPU | Squares: iGPU | Diamond: cloud



* Gemma NPU: 9/10, diagnostic. gpt-oss: 102.5 s from one successful request, omitted. gpt-6-luna: no valid speed result.

© 2026 Metrum AI, Inc.

Tools, JSON and code were near perfect; math separated models.

Fixed core suite; subset denominators are shown explicitly. These small samples cannot name a statistical winner.

Model / backend	Math	Instructions	JSON	Tools	Code	Summary	Overall
Qwen3.6-35B · NPU	14/15	14/15	8/8	8/8	8/8	6/6	58/60
Qwen3.6-35B · iGPU	12/15	15/15	8/8	8/8	8/8	6/6	57/60
Gemma 4 12B · NPU	12/15	14/15	8/8	8/8	8/8	6/6	56/60
Gemma 4 12B · iGPU	13/15	14/15	1/8	8/8	8/8	6/6	50/60
Qwen3.5-9B · NPU	11/15	13/15	8/8	8/8	8/8	5/6	53/60
gpt-4o-mini	12/15	13/15	8/8	8/8	8/8	6/6	55/60
gpt-6-luna	11/15	14/15	8/8	8/8	8/8	6/6	55/60
gpt-oss-20B · NPU · 26-case subset	6/7	8/8	4/4	n/a	4/4	2/3	24/26
Qwen3.8-27B · NPU · 30-case subset	7/7	7/8	4/4	2 pass, 2 ungraded	4/4	3/3	27/28 graded

Gemma iGPU: 7 JSON answers fenced, no credit. gpt-oss tools: n/a (not advertised by FastFlowLM). Qwen3.8: 30-case subset, 27/28 graded; two tool calls ungraded.

© 2026 Metrum AI, Inc.

Same model, two engines: compare complete serving stacks.

Model / backend	TTFT p50	Decode tok/s	Core	Math +30	MMLU-Pro	Retrieval	Extended	Peak process memory
Qwen3.6-35B · NPU	16.19 s	13.49	58/60	27/30	7/10	6/6	40/46	25.5 GiB
Qwen3.6-35B · iGPU	5.52 s	19.77	57/60	30/30	7/10	6/6	43/46	31.1 GiB
Gemma 4 12B · NPU	6.16 s*	4.99*	56/60	29/30	9/10	6/6	44/46	~10.5 GiB
Gemma 4 12B · iGPU	12.94 s	9.89	50/60	29/30	9/10	6/6	44/46	20.4 GiB

Speed

iGPU decode: about 1.5–2× faster. NPU first token is faster for Gemma, slower for Qwen.

Quality

No backend wins everywhere. All retrieval runs passed through a 16k-token window.

Memory

NPU working sets were smaller. Not total device memory; context settings differed.

* Gemma NPU speed: 9/10, diagnostic. Different quantization and exports: not a silicon-only comparison.

© 2026 Metrum AI, Inc.

Whisper on the NPU beat the CPU on these clips.

Three public LibriSpeech test-clean clips: 2.7 s, 7.8 s and 1.8 s; English; CC BY 4.0.

Backend	Runtime / precision	OK	Latency p50 / p95	Real-time factor	WER / CER	Device proof
NPU	FastFlowLM 1.0.10 · Whisper-v3 turbo Q4_1	10/10	2.74 / 3.16 s	0.95	0 / 0	NPU busy 18.3 s of 34.5 s · expected process · iGPU 0%
CPU	faster-whisper 1.2.1 · large-v3 turbo int8, 12 threads	10/10	4.33 / 4.44 s	1.52	0 / 0	NPU 0 · iGPU 0%

Pilot-ready, within the tested scope

10/10 per backend; WER and CER 0 / 0. Broader accuracy testing is still needed.

Keep the comparison honest

Latency includes HTTP and transcript capture. Different precision and decoders; only three clean clips.

CPU fix: bypass incompatible PyAV path and read 16 kHz WAV directly, with no resampling. NPU results were not rerun.

© 2026 Metrum AI, Inc.

How the night went: platform to scope decisions.

Time → problem / cause → fix or decision. Chronological order; corrected report chapter labels retained.

1 · Platform unblock

5:46 PM	Restore disabled on C: → enable; point verified 5:47 PM.
5:49 PM	Stale System32 xrt-smi → driver-folder copy; no downgrade.
5:52 PM	SDK 1.8 resets/timeouts → mark blocked; use other runtimes.
5:52 PM	NPU ready → Llama 1B runs; owner flm, iGPU 0%.
6:30 PM	Linux client / Windows server → WSL mirrored localhost.
6:32 PM	4k prompt mix unavailable → fixed, hashed exploratory prompts.

2 · Trustworthy numbers

7:32 PM	“Thinking off” overridden → reasoning_effort:none + warm-up check.
7:35 PM	Eval syntax/path errors → expressions + relative path; preflight passed.
7:41 PM	Qwen stops early; ignore_eos ignored → length FAIL; actual tokens.
8:14 PM	Grader settings / missing grade → fix; rerun affected cases only.
8:24 PM	gpt-6-luna HTTP 400; max_tokens sent → file #357; no speed.
8:28 PM	Evidence privacy → redact cookies/IDs; key scans clean.

3 · Large-model time limits

8:47 PM	27B under 1 tok/s; cause unproven → stop; fixed subset.
9:22 PM	Gemma metadata type error → correct; speed 9/10 diagnostic.
9:34 PM	Controller restarted → one job at a time; verify idle between.

4 · Scope decisions

9:48 PM	Operator keeps top five LLMs → smaller models cut; router skipped.
10:02 PM	Operator cuts other modalities and power → no active job interrupted.
10:27 PM	Selector misses valid mix with extra rows → offline repro; file #358.

© 2026 Metrum AI, Inc.

How the night went: recovery to evidence freeze.

Time → problem / cause → fix or decision. Chronological order; corrected report chapter labels retained.

5 · NPU failure and recovery

10:30 PM	gpt-oss: 9/10 hit reasoning cap → separate thinking-required row.
10:55 PM	Impossible busy share; wrong window → matched windows; 97.0%.
11:08 PM	Connection limit reached (10) → iGPU work; restart, then rerun NPU.

6 · iGPU + recovered NPU runs

11:11 PM	Qwen iGPU: pinned GGUF → 10/10, exact length, 57/60.
12:04 AM	Gemma iGPU: fenced JSON → strict 50/60 retained.
12:34 AM	Qwen 9B recovered → 10/10; overwritten startup metadata disclosed.

7 · Score correction / 6b · NPU extended

1:19 AM	showThinking polluted checks → match saved final answers; re-grade only. Corrected 11/26 invalid diagnostic → 24/26; no new model calls.
1:34-2:01 AM	NPU extended reruns → Qwen 40/46; Gemma 44/46; retrieval passed.

8 · Speech-to-text and wrap-up

2:02 AM	NPU Whisper → 10/10, 2.74 s, no transcription errors.
2:03 AM	CPU PyAV mismatch → direct WAV read; 10/10, 4.33 s.
2:20 AM	Evidence frozen → 1,071 files verified; sampler stopped 2:23 AM.
2:25 AM	Evidence and history → restic snapshot 34754521.

© 2026 Metrum AI, Inc.

Issues have owners; the open serving gaps are explicit.

Issue	Owner	Status	Effect here
Can't omit max_tokens for models that reject it	bench-cli	bench-cli #357 Filed / triaged; fix PR awaits approval	No gpt-6-luna speed numbers
Prompt selector misses a valid mix when non-matching rows are present	bench-cli	bench-cli #358 Filed	Built fixed-length prompts instead
Server stops accepting connections after 10 are left open	FastFlowLM 1.0.10	Not filed upstream yet	Several NPU jobs failed until a restart
ignore_eos / min_tokens not honored	FastFlowLM	Not filed	NPU rows fail the exact-length check
reasoning_effort overrides think:false	FastFlowLM (documented behavior)	Worked around	Needs reasoning_effort:none
Stale xrt-smi in System32	AMD driver install	Worked around	False "driver too old" signal in Phase 1
Ryzen AI 1.8 download host resets connections	AMD	Unresolved	Ryzen AI SDK 1.8 not tested
showThinking on by default	promptfoo (our configuration)	Fixed and re-graded	gpt-oss score corrected from 11/26 to 24/26
faster-whisper 1.2.1 incompatible with PyAV 19	Upstream libraries	Worked around	One failed CPU warm-up
Wrong value types in machine metadata, and a busy-share calculation error	Our harness	Fixed	No effect on final numbers

© 2026 Metrum AI, Inc.

Reported costs only. Power and energy were not measured.

Item	USD	Basis
Cloud reference answers and grading (gpt-4o-mini, gpt-6-luna, gpt-6.1-sol grader)	0.068	Token usage returned by the API, at list prices. Not reconciled with the invoice
pi agent that ran Phase 2 (gpt-6.1-sol, high effort, 8.6 hours)	22.91	Token usage reported by the agent, at list prices
Phase 1 router cloud calls	0.0024	Router usage rows
Local hardware, electricity	–	Company laptop. Not metered
Evidence storage (restic, 7.4 MiB)	≈ 0	Shared S3 artifact repository

No energy claim

Windows device counters read 0; package counter units unvalidated. Hardware and electricity not metered.

Not tested

Ryzen AI SDK 1.8; embeddings, image, vision, VLMs; router integration; concurrency beyond 4.

© 2026 Metrum AI, Inc.

The method separates performance, quality and device proof.

Speed

bench-cli 1.6.0 in WSL; Windows localhost; warm-up, streaming, ~1,000-token prompts, target 256 tokens; 10 serial requests.

Quality

promptfoo 0.125.1; 60 fixed cases, temperature 0. Summaries graded by gpt-6.1-sol; generated code in a network-less sandbox.

NPU proof

DXCore adapter ID; cumulative Running Time about every 1.15 s, with process owner and iGPU utilization. No NPU row had zero busy time.

Evidence snapshot 34754521

1,071 Phase 2 files: scripts, prompts, per-request results, eval outputs, transcripts and device proofs. Full study history and raw sampler CSVs are also archived.

```
metrum-artifacts restore 34754521 <dir>
```


© 2026 Metrum AI, Inc.

Phase 1 recap: the NPU did the prefill.

R^2 0.991

NPU busy time tracks prompt length. It stays flat as generated output grows; non-NPU paths read zero.

Edge-first works with workarounds

Plain text runs locally. Tool calls fall back to cloud. Schema and Anthropic-bridge traffic need eligible cloud targets. Native Windows router works; Windows tests and CI still need fixes.

No showstopper found.

Hybrid inference: NPU prefill → iGPU decode. This is the Phase 1 result, not a Phase 2 router retest.

© 2026 Metrum AI, Inc.

Current models run locally. Route by the workload.

Qwen3.6-35B-A3B is the practical NPU LLM here.
The same model is faster on the iGPU.
Whisper on the NPU is ready for a pilot.

Keep the limits visible

Supervise the server. Preserve strict quality scores. Do not claim power savings without measuring power.

Full report and evidence available from Metrum AI

© 2026 Metrum AI, Inc.