Poster

Live Demo
6 ARC-AGI scenarios

AGENTIC AI AT THE EDGE

Cost-Aware Orchestrator & Benchmarks

Kulbir Singh Ahluwalia
University of Illinois Urbana-Champaign
PhD Candidate, Computer Science
Team: Compute BU
Mentor: Alok Sharma
Manager: Mandar Deshpande

DEMAND FOR ACCELERATED COMPUTE IS GROWING FASTER THAN SUPPLY: AND THE GAP IS INCREASING.

Agentic AI runs everywhere: a Cost-Aware Orchestrator decides *WHERE? & HOW?* per task.
Route each agentic task to the cheapest *capable* device across **8 Compute devices**, **NPU / GPU** accelerators, ARM64 & AMD64 hosts, **Multiple OS deployments**

Goal: Minimize TCO(Total Cost of Operation) & Latency. **Result: 51%** of agentic commands served on an **18.3 W** edge NPU at comparable tool-calling accuracy, **2.4×** fewer GPU tokens.

1 Problem Formulation: Edge Agentic Deployments

- **Agentic AI exponentially increases Token demand:** Total inference cost, not quality, decides placement:
- **Inference Optimization:** Hardware differences in bandwidth & power decide Total Cost of Operation (TCO).
- **Current Orchestration tools optimise for resource utilisation:**, not measured joules per task.
- **Cost Aware Routing:** Power efficient silicon supplements high end accelerators.

Takeaway: Optimize for multi device, multi architecture, multi OS cost aware agentic deployments.

2 Related Work: SOTA Edge, Hybrid, Cloud KPIs

Device	Tok/s	TTFT (s)	Tool Sel	BW (GB/s)	Power (W)
IQ9 Hexagon NPU	18.3±0.1	0.19	**1.00**	37[d]	**18.3**[D]
GB10 DGX Spark	76.1[†]	0.043	**1.00**	140	51.6[R]
Jetson Thor	51.3±1.2	0.075	0.59[m]	163.8	65.9[B]
RTX 5090	**327**±0.5	**0.017**	0.96[m]	**1069**[s]	285.6[R]
Jetson Orin 64 GB	38.1±0.05[o]	1.64[o]	**1.00**[q]	n.m.	49.2±2.2[M,o]
GH200 ×4 (cloud)[‡]	43	n.m.	0.89	3608[s]	n.m.
H200 141 GB (ICRN)[h]	107.8[h]	0.44[h]	**1.00**[h]	n.m.	174[R]
H200 143 GB (ICC)[H]	n.m.	n.m.	**1.00**[q]	n.m.	n.m.

Takeaway: IQ9 competes closely with GPU's tool-selection accuracy at 1/15th power.

3 Quantitative Evaluation: Edge Compute & Use Cases

Use case measured gate · SLO	IQ9 Hexagon NPU · 36 GB 18.3 W[D]	Jetson Thor unified · 122 GB 65.9 W[B]	GB10 Spark unified · 128 GB 51.6 W[R]	RTX 5090 discrete · 32 GB 285.6 W[R]	A100-80GB ICC SECONDARY not metered	H200-141GB ICC SCAVENGER not metered
Agentic tool-calling 27-case hard suite gate sel & arg ≥0.90 · SLO ≤10 s	**n.m.**[1] · **n.m.**[1] 3.0 GB[4] · 6.38 s 1.00/1.00 · qwen3-4b	**n.m.**[1] · **86 ms** n.m.[2] · 0.76 s 1.00/1.00 · 4b-2507	**n.m.**[1] · **54 ms** n.m.[2] · 0.77 s 1.00/1.00 · qwen2.5:7b	**n.m.**[1] · **17 ms** n.m.[2] · 0.17 s 1.00/1.00 · qwen2.5:7b	**n.m.**[1] · **n.m.**[3] n.m.[3] · 0.42 s 1.00/1.00 · qwen2.5:7b	**n.m.**[1] · **n.m.**[3] n.m.[3] · 0.32 s 1.00/1.00 · qwen2.5:7b
Long-horizon planning SokoBench corridor sweep gate solved >0.25 @ ℓ=25 · SLO ≤120 s	**14.3 t/s · n.m.** 3.0 GB[4] · 132 s 2/8 @ℓ25 · qwen3-4b	**50.8 t/s · n.m.** n.m.[2] · 88 s 0/8 @ℓ25 · 4b q4	**60.7 t/s · n.m.** n.m.[2] · 187 s 6/8 @ℓ25 · 4b +**think**	**139.3 t/s · n.m.** n.m.[2] · 27 s 0/8 @ℓ25 · 4b q4	**n.m.** no SokoBench run	**n.m.** no SokoBench run
LLM text-gen summarise, ≥32-tok outputs gate decode ≥10 t/s · SLO TTFS ≤1.0 s	**18.3 t/s · 188 ms** 3.0 GB[4] · n.m. 128 tok · qwen3-4b	**50.8 t/s · 86 ms** n.m.[2] · n.m. ≥2k gen tok · 4b[5]	**69.8 t/s · 39 ms** n.m.[2] · n.m. ≥2k gen tok · 4b[5]	**139.3 t/s · 15 ms** n.m.[2] · n.m. ≥2k gen tok · 4b[5]	**n.m.** [3] no decode/TTFS	**n.m.** [3] no decode/TTFS
CV perception classify · detect · depth gate runs on target unit · SLO ≤33 ms	**n/a**[6] · **n/a**[6] 6.7 / 8.5 / 10.7 ms Inc/YOLO11/MiDaS	**n.m.** no CV benchmark	**n.m.** no CV benchmark	**n.m.** no CV benchmark	**n.m.** no CV benchmark	**n.m.** no CV benchmark

Cell slots: decode tok/s · **TTFS** first token | peak memory · total time | *gate value · model.* **Colour is mechanical, per (use case, model)**: **G1** model serves · **G2** the row's printed quality/throughput gate · **G3** its SLO; ■ optimal, all applicable gates pass · ■ one of G2/G3 fails · ■ G1 fails, both fail, or measured zero-success · ☐ n.m. asserts nothing.

1.00 vs 0.22 on one device: cells name models. **Deep KPI 1, thinking budget**: sole uncensored split, GB10 qwen3:4b +think: median **9,267** completion tokens on solves vs **11,066** on failures (n=40/63, cap 12,288): failure = budget *exhaustion*. Others are capped (IQ9 2,048; Thor / RTX 5090 / GH200 3,500): *censored, not measured.*

Deep KPI 2, peak KV cache (allocator-read, clusters only): **1,792 MiB** (qwen2.5:7b, 32,768 cells / 28 layers), **10,240 MiB** (qwen3:32b, 40,960 / 64), identical across A100-80GB/H200. IQ9 0.604 GB: *derived* (ctx 4096), not runtime peak. [1]tool-call outputs undercut the ≥32-output-token decode-validity floor: 5090 327 t/s short vs 139 t/s at ≥2k.

[2]whole-device memory sample, no per-model peak. [3]cluster sbatch: latency, accuracy only. [4]IQ9: per-process VmHWM load peak (non-VRAM boundary). [5]decode from ≥2k-token SokoBench; TTFS: warm prompt-eval median. [6]no token stream. [D]device meter · [B]board VIN · [R]GPU rail: **never ranked across boundaries**. SD1.5 (6.11 s/img, one device): footnoted.

Fig. 1: the 18.3 W Hexagon NPU is optimal in 3 of its 4 measured workloads: more than any other device here.

Takeaway: Only one device is yellow on planning, and it pays 187 s for it.

4 Method: Distributed Cost Aware Orchestration

Route every task to the device $d^\star$ in a fleet spanning mobile NPU → edge NPU → AI box → discrete GPU → cluster:

$$d^\star = \arg\min_d \left(w_1 \, \mathrm{cost}^{\$}_{\mathrm{energy}}(d) + w_2 \, \mathrm{latency}(d)/\mathrm{SLO} \right) \quad \text{s.t.} \quad \mathrm{mem}_{\mathrm{weights}} + \mathrm{mem}_{\mathrm{KV}} \le \mathrm{VRAM}(d)$$

In practice the cost term is min–max normalised within the admissible device set, so w_1, w_2 act as comparable weights.

- Owned hardware pays energy × site tariff; rented API pays $/Mtok; a grant cluster is $0.
- Decode is bandwidth-bound: tok/s ≈ BW ÷ bytes-per-token (**roofline self-check 1.1%**).
- Cost is **route-dependent**: 3.5× tariff spread ⇒ **lowest-power ≠ lowest-cost**.

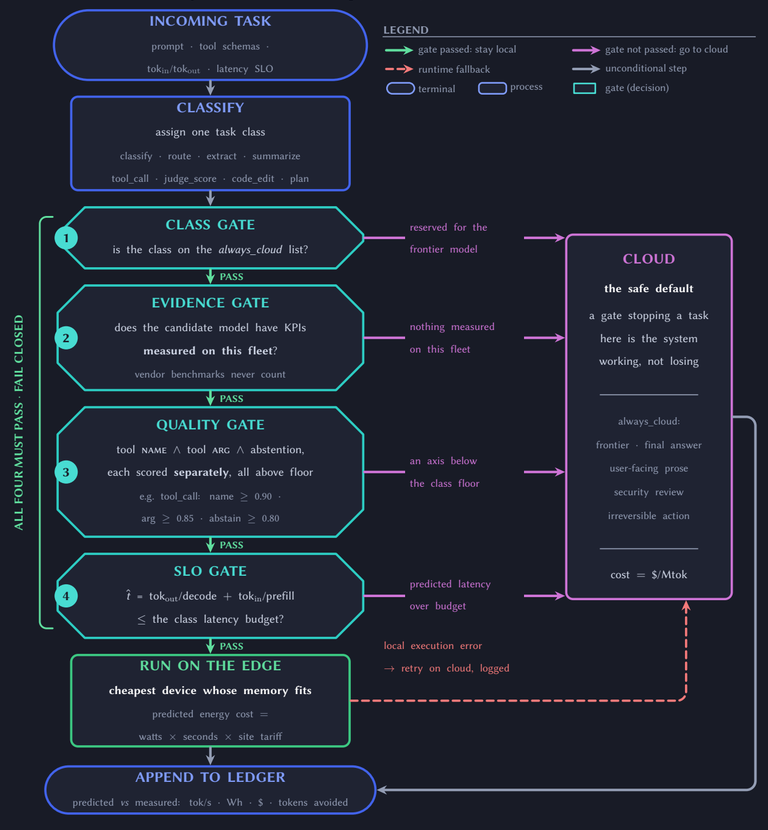


Fig. 2: all four gates must pass to stay local; any failure fails closed to the cloud rail.

Takeaway: Predict by roofline; price the route; place cheapest.

5 Model Deployment Benchmarks

Model-size confound removed: IQ9 NPU (18.3 W) vs RTX 5090 (286 W), median: it splits three ways:

WINS short/structured	**TIES** code-gen	**LOSES** long reasoning
json 1.6 s vs 8.7 s tool-use 2.3 s vs 7.6 s	12.4 s vs 12.2 s	gsm8k 27.2 s vs 10.8 s

Identical 24-case suite: IQ9 matches GB10 : tool-sel 1.00, semantic-arg 1.00, strict 0.917 on *both* (whitespace only). Median complete tool-call **3.72 s**; TTFT 188 ms is first-token only.

Use case	Decode[†] NT	Decode[†] Think	TTFT NT	TTFT Think	Peak mem NT	Peak mem Think
LLM text-gen · Qwen3-4B NPU	18.3	18.3	188 ms@227t	0.1 s	3.0 GiB[f]	3.0 GiB[f]
LLM text-gen · Qwen3-8B NPU[e]	11.2	10.9	0.1 s	0.1 s	5.2 GiB[f]	5.2 GiB[f]
VLM image-in · Qwen3-VL-4B NPU	16.8		509 ms@282t		2.2 GiB[f]	
Small LLM · 1.5B q4_0 CPU	43.4		1.00 s		1.9 GiB	
Small LLM · 1.5B q4_0 Adreno	42.6		1.03 s		1.9 GiB	
Image gen · SD1.5 NPU	6.11 s/img[b]		–		0.83 GiB	
CV classify · InceptionV3 NPU	5.9 ms/inf		–		44 MB	
CV depth · MiDaS v2 NPU	8.8 ms/inf		–		39 MB	
CV segment · DeepLabv3	*blocked: QAIRT 2.45/2.38 toolchain wall: needs off-board 2.38 re-export*					
CV detect · YOLOv11-640 NPU	8.5 ms/inf[g]		–		266 MB	
Agentic tool-call · 4B NPU	15.0	20.2	0.7 s[c]	*pending*	3.3 GiB[d]	3.0 GiB

Legend: measured · measured, caveat[b–d] · *pending / NOT MEASURED* · – = not applicable · NT = non-thinking · single-mode rows span both columns.

[†]decode tok/s (median); CV rows ms/inference; image-gen s/image. [b]derived from measured per-stage times (UNet 114.7 ms/step). [c]streamed first-token on an ~800-tok agentic prompt (tool schemas); complete-call median 3.72 s is separate. [d]geniex-serve campaign-level peak RSS, 4B resident. [e]largest shipped Genie bundle: runs on-board. [f]VmHWM load-peak host RAM; steady RSS ~0.16 GiB (weights live in fastrpc dma-buf). [g]100% HTP-delegated (QNN TFLite delegate); CPU 58.5 ms ⇒ 6.8×.

Takeaway: NPU **wins** short, **ties** code, **loses** long.

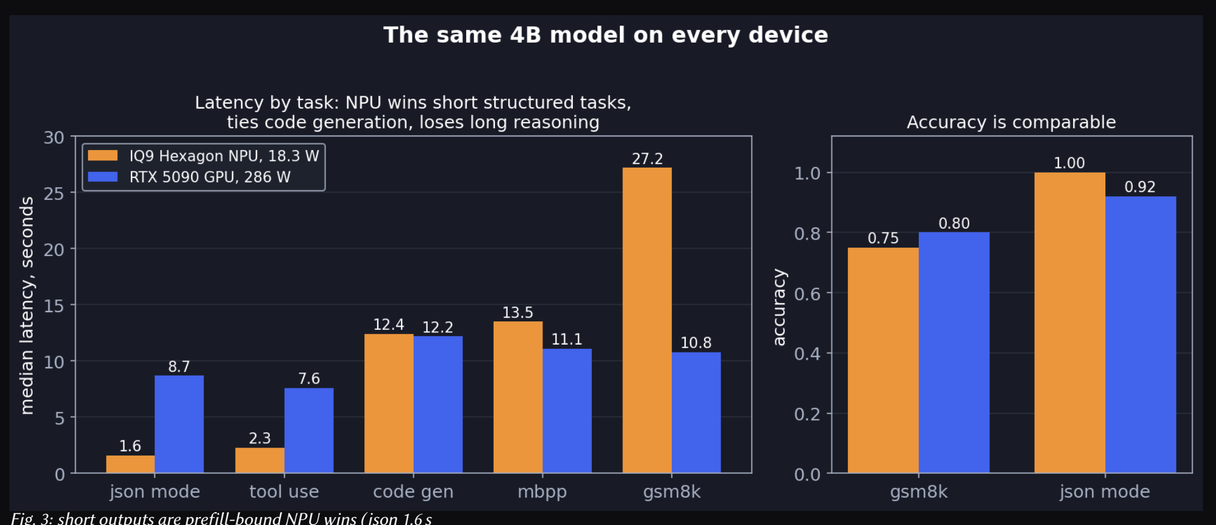


Fig. 3: short outputs are prefill-bound NPU wins (json 1.6 s vs 8.7 s); long outputs are decode-bound and bandwidth-limited GPU wins (gsm8k 27.2 s vs 10.8 s).

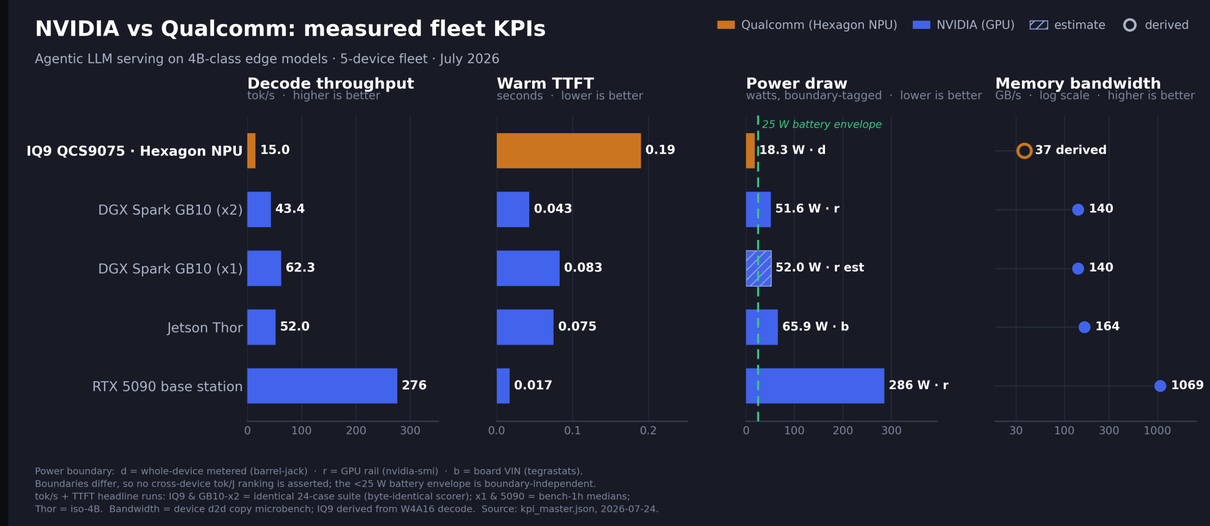
6 Edge Token Efficiency & Privacy

- **Data never leaves the device**: W4A16 on the Hexagon NPU (Genie/QNN, no CUDA, no cloud egress); **0.1 s** on-device TTFT, no queue variance. Inter-device context rides Tailscale WireGuard (Curve25519 + ChaCha20), secret-redacted before any write.
- **KG-at-edge token reduction**: session context as a knowledge-graph digest, not the raw transcript, compresses **689×** (14.7 MB → 21 KB, 0.24 s, $0, *zero* model tokens); content-addressed sharing sends only new bytes (≤2 KB/entry).
- **Cost-aware orchestrator** (40-KPI canon, K1–K40): a tool-call subtask = **$0.0105** (Claude Sonnet 5) vs **$0.000044** Urbana electricity on a GB10: **240×**; live demo savings **4.9–31×**. $/Mtok (list): Opus 5 5/25 · Sonnet 5 3/15 · GPT-4o 2.5/10.

Takeaway: Edge keeps data private and cuts context tokens 689× via KG representation.

Collaboration advantage: one orchestrator spans **4 OSes**, ARM & x86, NPU/GPU/edge/cloud: teams share compact knowledge-graph context, not raw data.

6 KPIs for Agentic Benchmarks

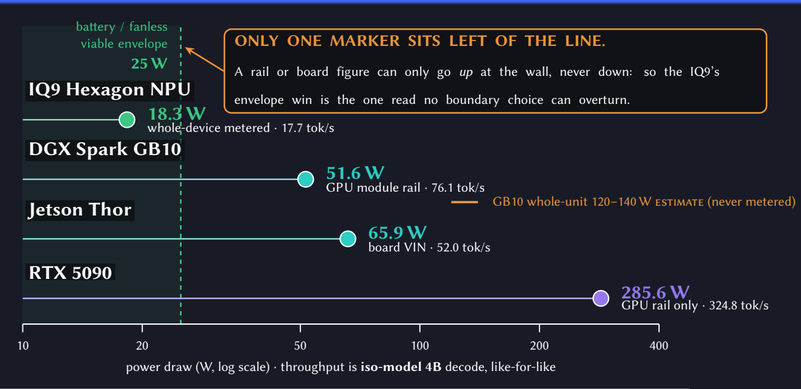


Power boundary: d = whole-device metered (barrel-jack) · r = GPU rail (nvidia-smi) · b = board VIN (tegrastats).
Boundaries differ, so no cross-device tok/J ranking is asserted; the <25 W battery envelope is boundary-independent.
tok/s + TTFT headline runs: IQ9 & GB10-x2 = identical 24-case suite (byte-identical scorer); x1 & 5090 = bench-1h medians; Thor = iso-4B. Bandwidth = device d2d copy microbench; IQ9 derived from W4A16 decode. Source: kpi_master.json, 2026-07-24.

Fig. 4: 4B-class agentic deployment. The IQ9 (18.3 W) is the **only** unit inside the 25 W envelope and **matches Nvidia DGX Spark GB10 GPU accuracy** at ~1/3 the decode rate.

Takeaway: GPUs win throughput; IQ9 matches accuracy under 25 W.

7 Total Cost of Operation: Inference & Network costs



Family (what the number contains)	watts	boundary / provenance
x86 CPU package (285HX; cores+uncore only: AMD not measured)	20.8–55.2	CPU-package / measured-RAPL
NVIDIA GPU rail (GB10/5080/5090/H200; silicon+VRM, no host)	15.3–285.6	GPU-rail / measured-rail
NVIDIA Jetson board (Thor VIN; SoC+board, no PSU)	16.3–65.9	board-VIN / measured-rail
Qualcomm QCS9075 CPU+Adreno+Hexagon (volts, no amps)	n/a	no rail exists / unavailable
Qualcomm IQ9: *everything* (+LPDDR5+UFS+PMIC+fan+IO)	**18.3**	whole-device / measured-external

Same MSI sweep, same instant: GPU rail 21.7 W + CPU pkg 20.8 W = 42.5 W: a *partial* sum (no RAM, fans, display, NVMe, PSU) vs the IQ9's **complete 18.3 W: 2.3×, a lower bound vs a total. Boundaries differ per row and are labelled per row**; ranges = [min..max] measured; no cross-boundary ratio or tok/J ranking is claimed.

Takeaway: A rail number is a floor; 18.3 W is a total. The win is the *envelope*, not tok/J.

8 Agentic Model Selection & Benchmarks

27-case hard suite, every model × (native + prompt-injection); best-mode tool-selection. Campaign shape & footnote keys: methodology strip (bottom).

Model	MoE	native	prompt
granite4:3b	dense	**1.00**	**1.00**
qwen2.5:7b	dense	**1.00**	**1.00**
qwen3.5:4b	dense	0.96	**1.00**
IQ9 qwen3-4b (NPU)[*]	dense	n/a	**1.00** (n=15)
Gemma 4 E4B	dense	0.89	0.93
gpt-oss:20b	MoE	0.89	**0.26**
granite3.1-moe:3b	MoE	0.15	**1.00**
Orin AGX 64GB qwen2.5:7b[‡]	dense	**1.00**	**1.00**
qwen3:32b (cloud)[c]	dense	0.93	0.89

Report best-of-{native, prompt}: invocation mode moves scores more than model choice (granite3.1-moe 0.15→1.00; gpt-oss:20b −0.63). *Why the MoEs lose: they route each token to a few active experts; the native tool-call template hits experts not tuned for structured output, so each MoE breaks in one mode (granite3.1-moe native 0.15, gpt-oss:20b prompt 0.26) while dense keeps all parameters on the format.*

Silicon-independent above the fit threshold: qwen2.5:7b holds 0.96:1.00 tool-selection across 7 tiers (Orin 64 GB, GB10, RTX 5080/5090, A100-40/80 GB, H200); median latency spans 0.21:1.86 s (8.9×). qwen3:32b: identical 0.93 native / 0.89 prompt *tool-selection* on A100-40/80 GB + H200 (tool-arg differs by 1 of 27). **Route by cost and latency, not GPU size.**

Takeaway: Robustness, not size, wins: dense 4–7B hold both invocation modes; the two MoEs each collapse in one.

10 Limitations & Conclusion

- **Open:** KV-cache vs context (unmeasured); full TCO (capex/ops uncosted); per-run NPU peak memory; GB10 Power estimates ⇒ no cross-boundary tok/J ranking.
- **Scope:** N=200, one task family; temp 0 ⇒ no error bars; single-tenant.

Conclusion: total inference cost, not quality, decided placement: solving $d^\star$ per task across a multi-OS, multi-architecture fleet turns edge silicon into dollars: $0.0005 metered vs $0.045 published-API per task (88×; 23–147× across providers), at tool-accuracy parity on 18.3 W.

11 Future Work: Fine-Tuned Orchestrator Models

- **Fine-tune dense + MoE orchestrators** (Qwen3-4B/8B dense; MoE: Qwen3-30B-A3B, gpt-oss-20B) to **predict all 14 KPIs** per (task, device, model) and emit the allocation $d^\star$ from the predicted KPIs; self-supervised labels from the fleet's measured bench corpus.
- **Add NPU speculative decoding**: 1.5–2.5× decode target [Li'24, Cai'24]
- **Quantize KV-cache to 2–4-bit** beyond Genie Q8_0 [Liu'24, Hooper'24]
- **Stream MoE experts from NVMe** (measured 3.1 GB/s tier) [Alizadeh'24]

18.3 W — Dragonwing IQ 9075 EVK Average Power Supply Reading

1.00 : 1.00 — Zero accuracy tax at the edge: NPU matches GPU, 24/24 both

51% — (Edge/Subnet+Cloud) Inference Ratio Secure, Lower Cost, Lower Latency

2.4× — Fewer GPU-tier tokens (435k vs 1.04M projected)

94.5 % — End-to-End validity (matches all-35B)

3.72 s — Median Tool Call Completion time

Methodology: **428,938** measured tool-calls, **138** device×model×mode cells (14 devices, 33 models, 27-case suite, temp 0), **Jul 20 – Aug 3 2026**; counted from valid summaries only. Temp-0 repeats = reproducibility; breadth = the 138 cells. Block-2 = iso-4B unless noted (W4A16 IQ9 / Q4 NVIDIA; GH200: GLM-5.2 744B MoE 2-bit). Not benchmarked (declared): MSI RTX 5080, MacBook M3 Max. Keys: [d]derived · [‡]spec · [m]bench-1h median (Thor 0.59 = thinking-4B; instruct-4B = 1.00) · [q]qwen2.5:7b · [h]ICRN 35B-A3B Q4, 12-case · [H]ICC full H200, qwen2.5:7b, 2,979 valid summaries (80,433 pooled); accuracy + 0.33 s median E2E only · [*]max across models / block-8: 15-case GenieX suite · [‡]Orin, accuracy only · [c]27-case suite on A100-40/80 GB + H200: identical tool-selection; latency 4.34 / 4.35 / 2.52 s · n.i. = not instrumented. Power: [D]whole-device meter · [o]Orin under co-tenant load (tok/s lower bound, rail power upper bound) · [M]Orin module-rail sum (INA3221) · [R]GPU rail · [B]board VIN: **no cross-boundary tok/J ranking**; derivations: repo README (QR).

Qualcomm