CLAIM REGISTER — AMD + Dell + Metrum AI Router launch media

Generated from **shared/CLAIMS.md**, the single source of truth for every number in `launch_deck.pptx`, `launch_overview.mp4`, `technical_story.mp4` and `internal_octopai.mp4`. If a number is not in this register, it does not go on screen. The deck carries the claim chips and the speaker notes; this register is the full manifest, shipped with the production files because the deck appendix is limited to three slides.

> Labels: **MEASURED (scope)** we ran it · **MODEL** arithmetic on stated assumptions · **VENDOR CLAIM** someone else's published figure, attributed · **ILLUSTRATION** drawn to explain, not a result · **METHOD** a structure, not a number · **PLAN** a proposed next step. S, B and V8 are source-deck labels, never claim IDs. Glyphs the brand subsets lack are transliterated (>=, <=, ~, x, ->, sum of).

1. Claims referenced by launch_deck.pptx (76)

C1	cost per accepted job = (inference + tools and sandbox + retries and failed attempts + verification + operating overhead) ÷ accepted jobs, with the latency requirement stated.	METHOD	PROMPT.md; COST-MODEL.md §1
R19	The quality floor is a prediction, not a guarantee.	FACT	S3; V8 51
A4	Failed vs passed SQLite attempts: median 48 vs 47 core-seconds (p10–p90 44–103 vs 42–112). Failures cost real CPU.	MEASURED (scope)	pilot #outcomes
C11	Cost receipt, frontier-priced: one accepted job = $0.0359 (inference $0.0326 + tools $0.0024 + retries $0.0008).	MODEL	model §4.4
C13	Retry lever: a failure rate of 20% / 33% / 50% / 67% multiplies the cost per accepted job by 1.25x / 1.49x / 2.00x / 3.03x.	MODEL	model §4.5; A14
A1	Per-agent CPU demand: SQLite 0.276 cores; CompCert 0.543 cores	MEASURED	pilot #formulas; seller-deck 4;

0	(stated in the source decks as "about 0.28 of a core per coding agent on EPYC").	(scope)	update-v8 27; battlecard 1
A14	Verified success: 4 of 5 low-load trials passed the verifier; at >=12 agents the pass rate ranged 33–69%; the heaviest loads passed about a third.	MEASURED (scope)	pilot #outcomes
I2	~250 concurrent coding agents per MI350P (Metrum tests: terminal-bench coding workloads + raw MI350P inference).	MEASURED (scope)	EVIDENCE-SEED I2
A5	Zero 429s does not mean no model-response waiting. 0 of ~4,000 calls throttled; max relay admission wait 0.0002 s; agents still spent a large share of each run waiting on the model.	MEASURED (scope)	pilot #model, #results (V4)
A8	Valid whole waves above 80% busy: 5 and 8 agents on 1 core (81.9%, 83.9%); 12 agents on 2 cores (82.8%). No valid 4-core whole wave exceeded 80% (highest load tested: 24 agents).	MEASURED (scope)	pilot #cpuonly
A11	Latency-safe load (secondary): estimate 2 / 5 / 9 agents on 1 / 2 / 4 cores; largest valid single waves that met the target held 3 / 7 / 14 agents at 67% / 72% / 57% busy.	MEASURED (scope)	pilot #sc-latency
A12	Registered verdicts: V1 FAIL (no load met >=90% busy with >=80% on time); V2, V3, V5 INCONCLUSIVE; V4 PASS (no throttling); V6 FAIL; V6′ FAIL.	MEASURED (scope)	pilot #results
A7	T1 = 184.7 s (median of the first three single-agent trials); SLO = 2 x T1 = 369.3 s (6.2 min).	MEASURED (sco	pilot glossary

		pe)	
A1	Scope: SQLite coding agents on pinned AMD EPYC 9R14 cores on AWS, Qwen3 Coder Next on Amazon Bedrock, 25–26 Oct 2026. It is not NPU, Halo or MI350P validation and does not establish Router uplift.	MEASURED (scope)	pilot #cpuonly, #model
A2	Primary sizing question: CPU-only — the number of agents that keeps >80% of the allocated CPU busy, regardless of completion or verifier result. Latency-safe load and verified success are separate measures.	MEASURED (scope)	pilot #cpuonly
A3	The 3 / 7 / 14 on-time loads were single waves, not validated sustained capacity. All six sustained studies were invalid (instrumentation faults); the held-out four-core prediction failed.	MEASURED (scope)	pilot #cpuonly, #holdout
A9	Provisional per-agent estimate: 3 / 6 / 12 agents at >80% busy on 1 / 2 / 4 cores, ranges 2–5 / 4–9 / 7–17.	MODEL	pilot #cpuonly
A15	Sealed 4-core prediction (SHA-256 b95bbabb…eeeb81, recorded before any 4-core run) expected 62–67% on time at 4–14 agents; measured 83–100%. Scored as registered: V6 FAIL.	MEASURED (scope)	pilot #holdout
R2	vLLM Semantic Router d3 live run (11 Oct): 204/204 test prompts routed to the intended tier, 104 ms p50, fail-closed verified; shipped in v4.2.	MEASURED (scope)	EVIDENCE-SEED R2
R3	OpenJev 93% classification on a 60-prompt test set, 193 ms	MEASURED	EVIDENCE-SEED R3

	decision latency.	(scope)	
R4	85% lower model spend vs the frontier API bill in one deployment (~25 developers). The demo's 92.5% is vs GPT-6 Astra list for one team.	MEASURED (scope)	EVIDENCE-SEED R4
R5	Per-agent key revocation is effective on the next request (~2 s across replicas). Key revocation stops model access; it is not whole-agent containment.	MEASURED (scope)	EVIDENCE-SEED R5
R14	The router pod is CPU-only and requests no GPU.	FACT	S5; V8 38
R15	100% of routed tokens land on a usage row; three ceilings per key (daily, monthly, lifetime); budgets reserve input plus the output cap before dispatch, then reconcile.	MEASURED (scope)	V8 18
R23	Neither decision model (d3, OpenJev) has been measured on Instinct yet; both ran on one NVIDIA RTX PRO 6000.	MEASURED (scope)	S13, S14; V8 47
R2b	d3 boundary: the live run was d3 26B in enforce mode, fail-closed verified when d3 was unreachable. It does not establish typical savings (1.4% on the seed set, 25% on one 8-turn take) or long-context latency.	MEASURED (scope)	V8 47; S13
R3b	OpenJev boundary: live run 25 Sep, 60 prompts x 2 on one RTX PRO 6000, 0 of 60 repeat flips; spend projected from a 15-call price sample. It does not establish production accuracy, Instinct performance, or commercial	MEASURED (scope)	V8 47; S13

	rights.		
R-G	Guardrail (not a claim): the Router is a gateway, not the serving runtime. No guaranteed quality, no aggregate atomic dollar budgets, no automatic privacy from a classifier, no whole-agent containment from key revocation. Verify release/runtime/model rights and distinguish vendor claims, measured results, models and illustrations.	FACT	PROMPT.md "EVIDENCE THAT MUST SURVIVE"; EVIDENCE-SEED
N1	Qwen3.6-35B-A3B on the NPU scored 58/60 on a fixed 60-case suite (cloud gpt-4o-mini: 55/60). Decode 13.5 tok/s on the NPU vs 19.8 on the iGPU (llama.cpp Vulkan).	MEASURED (scope)	NPU PDF "Same models on two chips"
N2a	A d3:nano decision took 0.40 s alone on the NPU.	MEASURED (scope)	NPU PDF "Decisions while a chat answer streams"
N2b	The same decision took 39 s when a chat model ran on the same NPU (the NPU serves one request at a time).	MEASURED (scope)	NPU PDF
N2c	With chat on the iGPU and decisions on the NPU, the decision took 0.45 s, all 50 decisions were correct, and chat was 9% slower.	MEASURED (scope)	NPU PDF
N4	NPU 93% busy while running a 35B chat model, CPU 4%, iGPU 0% (generation run).	MEASURED (scope)	NPU PDF "Which chip does the work"
N5	Whisper large-v3 turbo: 2.4 s/clip on the NPU vs 4.2 s on the CPU.	MEASURED (sco	NPU PDF

		pe)	
N6	FastFlowLM NPU kernels: free for commercial use and redistribution in unmodified form under its binary license (the orchestration code is MIT); upstream issues were filed (FastFlowLM #880–882, #680; RyzenAI-SW #415). Verify the current terms before shipping.	FACT	FastFlowLM README (FastFlowLM NPU Binary License) — supersedes the earlier "non-commercial" reading; NPU PDF "Issues we filed upstream"
H1	53 tok/s for one user on Qwen3.6-35B-A3B (llama.cpp Q4_K_M); 16 chat streams at 15.8 tok/s each on Gemma 4 E4B (253 tok/s total, 0.86 s first token, official rocm/vllm).	MEASURED (scope)	EVIDENCE-SEED H1
H2	Quality equal to NVIDIA GB10 within noise on the shared hard-tier suites.	MEASURED (scope)	EVIDENCE-SEED H2
H3	The Halo box is a Ryzen AI Halo device, not a Dell SKU — Dell has no AMD Halo product today; the 128 GB Dell Pro Max is GB10. AMD list price $3,999.	FACT	EVIDENCE-SEED H; blueprint §07
H5	GB10 is about 1.3x Halo for one user and level at 16 users (our data).	MEASURED (scope)	blueprint §07 ("our data")
P8	16 interactive users per Halo box; 13.5 tok/s background generation on the NPU.	MEASURED (scope)	blueprint §05b
I1	~7,800 output tok/s per MI350P on a 35B FP8 model (128->128, concurrency >=1024); 282 W GPU draw.	MEASURED (scope)	EVIDENCE-SEED I1
I2	3,150 output tok/s across 250 agents (~12.6 tok/s per agent); 282	MEASURED	blueprint §05b

b	W per GPU.	(scope)	capacity panel
I2c	~2,000 concurrent coding agents per 8-card XE7745 node = 8 x the tested ~250 per MI350P.	MODEL	S11; V8 25, 26, 35; B2
I2d	Sizing ladder: ~500 agents (R7725 + 2x MI350P, 288 GB HBM3E) · ~2,000 (R7725 + 1x XE7745, 8x MI350P) · ~4,000 (R7725 + 2x XE7745, 16x MI350P).	MODEL	S11; V8 26
I3	MI325X on DigitalOcean was a rented stand-in for the fine-tuning and demo runs. It is not MI350P performance and not on-premises proof.	FACT	EVIDENCE-SEED I3
I4b	XE7745: 8x MI350P, 1,152 GB HBM3E total, 4U, air cooled, 4.8 kW GPU, 8 PCIe GPUs, 2x EPYC 9005. MI350P: 144 GB HBM3E, CDNA 4, 128 CU, 4 TB/s, PCIe Gen5 x16, 600 W.	VENDOR CLAIM	S8; V8 46, 53
I4c	Dell's AI Platform with AMD names XE7745 and R7725 with MI350P and EPYC. The Dell catalog listing is in discussion, not live.	VENDOR CLAIM	S8; V8 22, 46; B2
F1	Run 1: tool-call strict exact match 4.5% -> 32.0%, but MMLU-Pro collapsed 60.5 -> 13.0 (IFEval 81.5 -> 63.5). The preregistered gate refused promotion.	MEASURED (scope)	EVIDENCE-SEED F1; inputs/fine-tuning-report.md
F2	The same frozen router export (1,186 / 221 / 300 traces) trained on MI325X and RTX PRO 6000 came out statistically level (semantic 78 vs 74/100); both run-2 candidates stayed in canary.	MEASURED (scope)	EVIDENCE-SEED F2; report
F4b	One audited export contained 1,707 traces (1,186 train / 221 validation / 300 held-out); the uncaptured	MEASURED (sco	S13; V8 13

	team was verified absent.	pe)	
F5	Serving cost per harness success: $0.00142 (AMD MI325X) vs $0.00130 (NVIDIA RTX PRO 6000) on one replay harness, 300 held-out traces, serving only, excludes training. Training wall time 893 s (NVIDIA) vs 1,215 s (AMD) — the decks round the AMD run to "about 20 minutes".	MEASURED (scope)	report; seller-deck 13; update-v8 14
F7	Run-2 gates: NVIDIA exact 5/200 vs base 4/200 (+0.5 pp); AMD 7/200 vs base 9/200 (-1.0 pp); +10 pp required.	MEASURED (scope)	report
F9	Teacher licensing: the provider's terms (§2.2(c)) prohibit using content to train AI models; real-teacher collection was stopped until a licensed teacher was chosen.	FACT	report
F10	0 of 2 candidates were promoted: the loop is proven, a better model is not yet.	MEASURED (scope)	S15; V8 13; B2
F11	LoRA training on governed, router-exported traces took about 20 minutes per run on Instinct MI325X (1,215 s; the NVIDIA run was 893 s).	MEASURED (scope)	S14; V8 14
C2	Illustrative workload (blueprint reference): 500 seats · 4 always-on agents per seat (2,000 agents) · 17 h/day · 30 days · 3% escalation to frontier · 12.6 tok/s sustained per agent · 46.3B agent output tokens/month · 12.1M jobs/month · ~3,704 output + ~74,088 input tokens per accepted job · node occupancy 68.7%.	ILLUSTRATION	COST-MODEL.md §3

C16	Price and cost inputs (from BLUEPRINT-REFERENCE.py): node capex $500,000 budgetary — not a quote; Halo box $3,999 x 32 boxes (1 per 16 users) -> total capex $627,968; DO MI325X $3.80/GPU-hr with no provisioning factor; GPT-6 Terra list $1/$5 per M tokens; agent input:output 20:1; 90% of agent input cached at 10% of list; electricity $0.12/kWh, PUE 1.4; node power 8 x 282 W measured + 1.5 kW host; ops $2,000/server/month + $20/Halo box; amortisation 36 months; 250 agents per MI350P, 12.6 output tok/s per agent.	MODEL	model §2
C12	Cost receipt, owned-priced at 68.7% busy: $0.0036 marginal / $0.0057 all-in per accepted job — ~9.9x below the frontier-priced job. Tool, retry and verification terms are identical across both columns.	MODEL	model §4.4
C3	All-frontier: $419,451/month for the modelled workload at GPT-6 Terra list ($1/$5 per M, 20:1 agent input, 90% cached at 10%).	MODEL	model §4.1
C4	Hybrid on-premises: $33,111/month all-in = node $16,350 + 32 Halo boxes $4,475 + cloud burst $0 + 3% escalation $12,287. 92.1% below all-frontier.	MODEL	model §4.1
C5	All-cloud (elastic): $34,644/month (elastic MI325X GPUs at $3.80/GPU-hr + hosted open models at $0.95/M + escalation). 4.4% above hybrid.	MODEL	model §4.1
C6	Break-even occupancy vs elastic cloud GPUs: 74.7% (~75%) — the node's all-in monthly cost equals the cloud cost of the same agent-	MODEL	model §4.2

	hours.		
C7	Payback vs all-frontier: 1.6 months.	MODEL	model §4.2
C8	Payback vs all-cloud: 33.1 months.	MODEL	model §4.2
C9	On-prem cost per M output tokens at the modelled 68.7% occupancy: $0.055/M marginal (power + ops only) and $0.364/M all-in (including capex ÷ 36).	MODEL	model §4.3
C10	On-prem $/M output tokens: at 40% busy $0.094 marginal / $0.626 all-in; at 90% busy $0.042 / $0.278.	MODEL	model §4.3
C14	Escalation lever: 1% / 3% / 10% / 25% of output tokens to a frontier API -> hybrid monthly $24,920 / $33,111 / $61,780 / $123,212. The single biggest lever on the hybrid bill.	MODEL	model §4.5
C15	Cached-input lever: 0% / 50% / 90% / 99% cached input -> all-frontier monthly $1,168,980 / $752,575 / $419,451 / $344,498. Cached input dominates the frontier price.	MODEL	model §4.5; G7
G7	Cache reads are 96–99% of agentic coding token spend, so any savings claim must use cached-input prices.	EXTERNAL (cited)	blueprint §07
A18	Reproduction: final snapshot restored to a fresh directory with every file hash matched; offline rebuild with network denied reproduced all 136 CSV files byte for byte; the sealed prediction hash was re-scored. Teardown verified 0 running instances, volumes, security groups, key pairs	MEASURED (scope)	pilot #data, #cost teardown

	or snapshots tagged for the pilot.		
A20	Follow-up work identified by the pilot: more matched repeats per load, a constant-backlog test to separate cohort drain from capacity, a second core type, and a measured 3-core point.	PLAN	pilot #followup
R17	The Apache-2.0 router has been in production internally since June 2026.	MEASURED (scope)	V8 40
R18	Mandatory disclaimer: testing performed by Metrum AI; performance and cost claims are Metrum AI's and have not been independently verified by AMD.	FACT	B2; V8 46
P9	Proposed next step for the enterprise tier: a Dell/AMD lab path to MI350P agent-workload validation, with no catalog or commercial commitment.	PLAN	PROMPT.md slide 12
P4	cost_per_accepted_job = sum of(attributable inference + tool/sandbox compute + retries and failed attempts + verification + operating overhead) ÷ verified completed jobs, with the latency requirement stated.	METHOD	PROMPT.md; blueprint §05
A6	$150.76 total ($59.19 compute + $91.57 model tokens) covers compute and model tokens across recovery, invalid and supplemental work, and excludes orchestrating-agent tokens. It is a study cost. Never divide it by SQLite successes.	MEASURED (scope)	pilot #cost

2. Full claim manifest (133)

N1	Qwen3.6-35B-A3B on the NPU scored 58/60 on a fixed 60-case suite (cloud gpt-4o-mini: 55/60). Decode 13.5 tok/s on the NPU vs 19.8 on the iGPU (llama.cpp Vulkan).	MEASURED (scope)	NPU PDF "Same models on two chips"
N2	Umbrella for the three decision latencies — always shown as three separate figures (N2a alone on the NPU, N2b contended, N2c split), never merged into one "NPU latency".	MEASURED (scope)	NPU PDF
N2a	A d3:nano decision took 0.40 s alone on the NPU.	MEASURED (scope)	NPU PDF "Decisions while a chat answer streams"
N2b	The same decision took 39 s when a chat model ran on the same NPU (the NPU serves one request at a time).	MEASURED (scope)	NPU PDF
N2c	With chat on the iGPU and decisions on the NPU, the decision took 0.45 s, all 50 decisions were correct, and chat was 9% slower.	MEASURED (scope)	NPU PDF
N3	d3:mini scored 24/24 on the labelled set.	MEASURED (scope)	NPU PDF
N4	NPU 93% busy while running a 35B chat model, CPU 4%, iGPU 0% (generation run).	MEASURED (scope)	NPU PDF "Which chip does the work"
N5	Whisper large-v3 turbo: 2.4 s/clip on the NPU vs 4.2 s on the CPU.	MEASURED (scope)	NPU PDF

N 6	FastFlowLM NPU kernels: free for commercial use and redistribution in unmodified form under its binary license (the orchestration code is MIT); upstream issues were filed (FastFlowLM #880–882, #680; RyzenAI-SW #415). Verify the current terms before shipping.	FACT	FastFlowLM README (FastFlowLM NPU Binary License) — supersedes the earlier "non-commercial" reading; NPU PDF "Issues we filed upstream"
H 1	53 tok/s for one user on Qwen3.6-35B-A3B (llama.cpp Q4_K_M); 16 chat streams at 15.8 tok/s each on Gemma 4 E4B (253 tok/s total, 0.86 s first token, official rocm/vllm).	MEASURED (scope)	EVIDENCE-SEED H1
H 2	Quality equal to NVIDIA GB10 within noise on the shared hard-tier suites.	MEASURED (scope)	EVIDENCE-SEED H2
H 3	The Halo box is a Ryzen AI Halo device, not a Dell SKU — Dell has no AMD Halo product today; the 128 GB Dell Pro Max is GB10. AMD list price $3,999.	FACT	EVIDENCE-SEED H; blueprint §07
H 5	GB10 is about 1.3x Halo for one user and level at 16 users (our data).	MEASURED (scope)	blueprint §07 ("our data")
I 1	~7,800 output tok/s per MI350P on a 35B FP8 model (128->128, concurrency >=1024); 282 W GPU draw.	MEASURED (scope)	EVIDENCE-SEED I1
I 2	~250 concurrent coding agents per MI350P (Metrum tests: terminal-bench coding workloads + raw MI350P inference).	MEASURED (scope)	EVIDENCE-SEED I2
I 2	3,150 output tok/s across 250 agents (~12.6 tok/s per agent); 282	MEASURED	blueprint §05b

b	W per GPU.	(scope)	capacity panel
I3	MI325X on DigitalOcean was a rented stand-in for the fine-tuning and demo runs. It is not MI350P performance and not on-premises proof.	FACT	EVIDENCE-SEED I3
I4	MI350P PCIe: 144 GB, air-cooled, up to 8 per server. Dell ships XE9785L (8x MI355X) and XE7745 / R7725 with MI350P.	VENDOR CLAIM	blueprint §07
I5	Shared KV cache on MI300X: 2.5–5.4x faster first token with LMCache/Mooncake on ROCm.	MEASURED (scope)	blueprint §06
I6	Dell lab health report (R7725, 2x MI350P x2): PSU input loss, PCIe errors, no OS login — a blocker for on-premises runs.	INT	blueprint §06
R1	Releases v4.0–v4.4 (21 Sep -> 27 Oct 2026), Apache-2.0; v4.4.0 ships Windows amd64/arm64 zips.	FACT	EVIDENCE-SEED R1
R2	vLLM Semantic Router d3 live run (11 Oct): 204/204 test prompts routed to the intended tier, 104 ms p50, fail-closed verified; shipped in v4.2.	MEASURED (scope)	EVIDENCE-SEED R2
R3	OpenJev 93% classification on a 60-prompt test set, 193 ms decision latency.	MEASURED (scope)	EVIDENCE-SEED R3
R4	85% lower model spend vs the frontier API bill in one deployment (~25 developers). The demo's 92.5% is vs GPT-6 Astra list for one team.	MEASURED (scope)	EVIDENCE-SEED R4

R5	Per-agent key revocation is effective on the next request (~2 s across replicas). Key revocation stops model access; it is not whole-agent containment.	MEASURED (scope)	EVIDENCE-SEED R5
R6	Tokenomics meter: reasoning + cache tokens, list vs billed cost, savings split by routing and caching.	FACT	EVIDENCE-SEED R6
R7	d3-ultra: open 27B multimodal decision model from AMD + the vLLM Semantic Router team, trained and served on Instinct, Apache-2.0, launched 26 Oct 2026.	VENDOR CLAIM	EVIDENCE-SEED R7
R11	Router built from source ran 90/90 against Lemonade.	MEASURED (scope)	blueprint §02b status table
R12	A sensitive verdict narrows the allow list to owned tiers, so the frontier group cannot be selected for that request whatever its difficulty.	FACT	blueprint §02b
R13	Local requests $0 vs $0.00223 per 30 cloud requests.	MEASURED (scope)	blueprint §05 ledger table
R-G	Guardrail (not a claim): the Router is a gateway, not the serving runtime. No guaranteed quality, no aggregate atomic dollar budgets, no automatic privacy from a classifier, no whole-agent containment from key revocation. Verify release/runtime/model rights and distinguish vendor claims, measured results, models	FACT	PROMPT.md "EVIDENCE THAT MUST SURVIVE"; EVIDENCE-SEED

	and illustrations.		
F1	Run 1: tool-call strict exact match 4.5% -> 32.0%, but MMLU-Pro collapsed 60.5 -> 13.0 (IFEval 81.5 -> 63.5). The preregistered gate refused promotion.	MEASURED (scope)	EVIDENCE-SEED F1; inputs/fine-tuning-report.md
F2	The same frozen router export (1,186 / 221 / 300 traces) trained on MI325X and RTX PRO 6000 came out statistically level (semantic 78 vs 74/100); both run-2 candidates stayed in canary.	MEASURED (scope)	EVIDENCE-SEED F2; report
F3	Run 1 trained on 1x NVIDIA RTX PRO 6000 Blackwell (Shadeform); GPU cost $3.19.	MEASURED (scope)	report §"run 1"
F4	Governed capture produced an audited export of 1,186 / 221 / 300 train / validation / held-out traces with the uncaptured team verified absent.	MEASURED (scope)	report; blueprint §04d
F5	Serving cost per harness success: $0.00142 (AMD MI325X) vs $0.00130 (NVIDIA RTX PRO 6000) on one replay harness, 300 held-out traces, serving only, excludes training. Training wall time 893 s (NVIDIA) vs 1,215 s (AMD) — the decks round the AMD run to "about 20 minutes".	MEASURED (scope)	report; seller-deck 13; update-v8 14
F6	Total measured fine-tuning spend ~$51 (GPUs ~$17, a short Fireworks deployment ~$32, teacher calls a few dollars).	MEASURED (scope)	report

F7	Run-2 gates: NVIDIA exact 5/200 vs base 4/200 (+0.5 pp); AMD 7/200 vs base 9/200 (-1.0 pp); +10 pp required.	MEASURED (scope)	report
F8	Router v4.3.0 validated and tagged 2026-10-24.	FACT	report
F9	Teacher licensing: the provider's terms (§2.2(c)) prohibit using content to train AI models; real-teacher collection was stopped until a licensed teacher was chosen.	FACT	report
A1	Scope: SQLite coding agents on pinned AMD EPYC 9R14 cores on AWS, Qwen3 Coder Next on Amazon Bedrock, 25–26 Oct 2026. It is not NPU, Halo or MI350P validation and does not establish Router uplift.	MEASURED (scope)	pilot #cpuonly, #model
A2	Primary sizing question: CPU-only — the number of agents that keeps >80% of the allocated CPU busy, regardless of completion or verifier result. Latency-safe load and verified success are separate measures.	MEASURED (scope)	pilot #cpuonly
A3	The 3 / 7 / 14 on-time loads were single waves, not validated sustained capacity. All six sustained studies were invalid (instrumentation faults); the held-out four-core prediction failed.	MEASURED (scope)	pilot #cpuonly, #holdout
A4	Failed vs passed SQLite attempts: median 48 vs 47 core-seconds (p10–p90 44–103 vs 42–112). Failures cost real CPU.	MEASURED (scope)	pilot #outcomes
A5	Zero 429s does not mean no model-response waiting. 0 of	MEASURED	pilot #model,

	~4,000 calls throttled; max relay admission wait 0.0002 s; agents still spent a large share of each run waiting on the model.	(scope)	#results (V4)
A6	$150.76 total ($59.19 compute + $91.57 model tokens) covers compute and model tokens across recovery, invalid and supplemental work, and excludes orchestrating-agent tokens. It is a study cost. Never divide it by SQLite successes.	MEASURED (scope)	pilot #cost
A7	T1 = 184.7 s (median of the first three single-agent trials); SLO = 2 x T1 = 369.3 s (6.2 min).	MEASURED (scope)	pilot glossary
A8	Valid whole waves above 80% busy: 5 and 8 agents on 1 core (81.9%, 83.9%); 12 agents on 2 cores (82.8%). No valid 4-core whole wave exceeded 80% (highest load tested: 24 agents).	MEASURED (scope)	pilot #cpuonly
A9	Provisional per-agent estimate: 3 / 6 / 12 agents at >80% busy on 1 / 2 / 4 cores, ranges 2–5 / 4–9 / 7–17.	MODEL	pilot #cpuonly
A10	Per-agent CPU demand: SQLite 0.276 cores; CompCert 0.543 cores (stated in the source decks as "about 0.28 of a core per coding agent on EPYC").	MEASURED (scope)	pilot #formulas; seller-deck 4; update-v8 27; battlecard 1
A11	Latency-safe load (secondary): estimate 2 / 5 / 9 agents on 1 / 2 / 4 cores; largest valid single waves that met the target held 3 / 7 / 14 agents at 67% / 72% / 57% busy.	MEASURED (scope)	pilot #sc-latency
A1	Registered verdicts: V1 FAIL (no load met >=90% busy with >=80%	MEASURED	pilot #results

2	on time); V2, V3, V5 INCONCLUSIVE; V4 PASS (no throttling); V6 FAIL; V6′ FAIL.	(scope)	
A13	Three repeats of 12 agents on 4 cores: 31% / 50% / 60% busy, 83% / 83% / 100% on time; one run used 54% more CPU work than another.	MEASURED (scope)	pilot #results
A14	Verified success: 4 of 5 low-load trials passed the verifier; at >=12 agents the pass rate ranged 33–69%; the heaviest loads passed about a third.	MEASURED (scope)	pilot #outcomes
A15	Sealed 4-core prediction (SHA-256 b95bbabb…eeeb81, recorded before any 4-core run) expected 62–67% on time at 4–14 agents; measured 83–100%. Scored as registered: V6 FAIL.	MEASURED (scope)	pilot #holdout
A16	Model speed: 1 request -> 0.79 s p50 time to first token, 4.3 ms per output token, 167 tok/s per request; at concurrency 8 -> 0.79 s / 5.0 ms / 137 tok/s; 20/0 accepted/throttled in the probe. Applied quota 10,000 requests/min, 100M tokens/min.	MEASURED (scope)	pilot #model
A17	Supplemental tasks (never pooled with SQLite): CompCert 573–1,984 core-seconds per attempt, 24–79% busy, 2 of 7 passed; pMARS 2–9% busy; POV-Ray 0–3% busy, 0 passed.	MEASURED (scope)	pilot #supplemental
A18	Reproduction: final snapshot restored to a fresh directory with every file hash matched; offline rebuild with network denied reproduced all 136 CSV files byte for byte; the sealed prediction hash was re-scored. Teardown	MEASURED (scope)	pilot #data, #cost teardown

	verified 0 running instances, volumes, security groups, key pairs or snapshots tagged for the pilot.		
A19	Qwen3 Coder Next was chosen from six Bedrock coding models; it solved 4 of 5 low-load trials (95% interval 38–96%); gpt-oss-120b solved 2 of 5 and was dropped.	MEASURED (scope)	pilot #model, journal
A20	Follow-up work identified by the pilot: more matched repeats per load, a constant-backlog test to separate cohort drain from capacity, a second core type, and a measured 3-core point.	PLAN	pilot #followup
P1	Working theme: "The token bill is not the agent bill." Earlier blueprint framing: "Right token, right silicon" — retained only as a closing line.	PLAN	PROMPT.md; blueprint §01
P2	Illustrative workload: one 8x MI350P server ~69% occupied; owned beats elastic cloud above ~75% occupancy; payback ~1.6 months vs all-frontier (budgetary capex).	MODEL	EVIDENCE-SEED P2; blueprint §05b
P3	OctopAI internal: 11 leads + 89 named subagents = up to 100 agents (concept/estimate); Claude coordinates. Real lessons: NPU contention (39 s vs 0.45 s); a fine-tune candidate rejected by its gate; claims discipline; the first preview film of this plan was made in ~50 minutes by one agent.	INT	EVIDENCE-SEED P3; blueprint §08c
P4	cost_per_accepted_job = sum of(attributable inference + tool/sandbox compute + retries and failed attempts + verification	METHOD	PROMPT.md; blueprint §05

	+ operating overhead) ÷ verified completed jobs, with the latency requirement stated.		
P5	NPU tier: $0 marginal, plus an optional device-share amortization; Dell's 5,000-seat model implies ~$50 per seat per year saved.	ILLUSTRATION	blueprint §05
P6	Strix Halo: $0.77–1.05 per M output tokens at 16 users and 25% busy. At 4 ops hours a month the box never beats a hosted API.	MODEL	blueprint §05
P7	XE7745 payback in about 10 weeks vs Opus at 90% cache hit.	MODEL	blueprint §05
P8	16 interactive users per Halo box; 13.5 tok/s background generation on the NPU.	MEASURED (scope)	blueprint §05b
P9	Proposed next step for the enterprise tier: a Dell/AMD lab path to MI350P agent-workload validation, with no catalog or commercial commitment.	PLAN	PROMPT.md slide 12
G1	AMD claims Helios delivers up to 30% more tokens per dollar than Vera Rubin NVL72.	VENDOR CLAIM	blueprint §07
G2	AMD reports that routing its own internal traffic to MI350P cut its token bill 43% (self-reported).	VENDOR CLAIM	blueprint §07b
G3	Dell's message: "tokenomics up to 90% below public cloud".	VENDOR CLAIM	blueprint §07
G4	Dell's deskside agentic AI offer (NVIDIA-only) claims up to 87%	VENDOR	blueprint §07b

	lower cloud token cost.	CLAIM	
G5	AMD claims up to 4.2x tokens per dollar for MI350P; MLPerf v6.1 debut.	VENDOR CLAIM	blueprint §07
G6	ROCm delivered +38% on the same MI355X between MLPerf rounds.	VENDOR CLAIM	blueprint §07
G7	Cache reads are 96–99% of agentic coding token spend, so any savings claim must use cached-input prices.	EXTERNAL (cited)	blueprint §07
G8	Goldman Sachs projects 24x token growth by 2030 from agents.	EXTERNAL (cited)	blueprint §07
G9	Forrester: open-weight models carry about half of tokens but ~5% of spend.	EXTERNAL (cited)	blueprint §07
G10	Dell's own survey: 59% cite unclear ROI on AI.	VENDOR CLAIM	blueprint §07
G11	Gartner: 40% of new AI data centers will be power-limited by 2027.	EXTERNAL (cited)	blueprint §07b
G12	Dell AI Factory: 6,500+ customers.	VENDOR CLAIM	blueprint §07
G13	Ryzen AI PRO 400 NPUs reach up to 60 TOPS; Qualcomm X2 Elite claims 80 TOPS; Intel Panther Lake	VENDOR CLAI	blueprint §07

	50 TOPS.	M	
G14	Schedule items: AMD/Dell follow-up 30 Oct; AMD Q3 earnings 3 Nov; CES 2027 6–9 Jan (AMD keynote unconfirmed); DTW 2027 24–27 May.	PLAN	blueprint §07b
G15	AMD is planned to lead the Router v5.0 launch; Dell AI Factory validation dated 11 Dec.	PLAN	blueprint §06
G16	Governance context: OWASP Agentic Top 10; NIST agent-identity work; EU AI Act high-risk obligations point to Dec 2027; ~1 in 5 organisations have mature agent governance (Deloitte, via Dell).	EXTERNAL (cited)	blueprint §07b
G17	Helios (MI455X, 72 GPUs per rack) is in production at OpenAI, Meta, Microsoft and Oracle; Helion is an OCP agentic rack with a built-in model router, partner configurations from November.	VENDOR CLAIM	blueprint §07
I2c	~2,000 concurrent coding agents per 8-card XE7745 node = 8 x the tested ~250 per MI350P.	MODEL	S11; V8 25, 26, 35; B2
I2d	Sizing ladder: ~500 agents (R7725 + 2x MI350P, 288 GB HBM3E) · ~2,000 (R7725 + 1x XE7745, 8x MI350P) · ~4,000 (R7725 + 2x XE7745, 16x MI350P).	MODEL	S11; V8 26
I4b	XE7745: 8x MI350P, 1,152 GB HBM3E total, 4U, air cooled, 4.8 kW GPU, 8 PCIe GPUs, 2x EPYC 9005. MI350P: 144 GB HBM3E, CDNA 4, 128 CU, 4 TB/s, PCIe Gen5 x16, 600 W.	VENDOR CLAIM	S8; V8 46, 53
I4c	Dell's AI Platform with AMD names XE7745 and R7725 with MI350P and EPYC. The Dell catalog listing	VENDOR CLAI	S8; V8 22, 46; B2

	is in discussion, not live.	M	
R 1 4	The router pod is CPU-only and requests no GPU.	FACT	S5; V8 38
P 7 b	MODELED: one XE7745 (8x MI350P) pays back in ~9 weeks against Claude Opus 5 list at 24x7 — ~$3.0M/yr of API spend vs a ~$500K budgetary node. Basis: 250 concurrent agents, ~103 generating at any instant at 30 output tok/s each, 3,150 output tok/s sustained, 90% cache hit, $5/$25 per M tokens. Without caching ~$11M/yr.	MOD EL	S7, S12; V8 28, 50; B2
P 1 0	MODELED: at 5,000 developers and public $250–1,500 per developer-month spend, the measured 67–85% cuts imply a $10–77M yearly reduction.	MOD EL	V8 2, 31
P 1 1	MODELED: at 5,000 developers and $250/$500 per developer-month, API spend is $15M/$30M a year and the router cuts $10–12.8M / $20–25.5M.	MOD EL	V8 31
P 1 2	The economic proof must separate three things and never count the same saving twice: route more efficiently · substitute local serving (include the complete local platform, operations and the API spend that remains) · utilization. Owned tokens are not free.	METH OD	V8 50

R15	100% of routed tokens land on a usage row; three ceilings per key (daily, monthly, lifetime); budgets reserve input plus the output cap before dispatch, then reconcile.	MEASURED (scope)	V8 18
R16	Salted PBKDF2-HMAC-SHA3-512 key verifiers (~4.7 us per cached request); upstream identity hidden from callers and error messages; credential values redacted from config, startup and CLI errors.	MEASURED (scope)	V8 20
R17	The Apache-2.0 router has been in production internally since June 2026.	MEASURED (scope)	V8 40
R18	Mandatory disclaimer: testing performed by Metrum AI; performance and cost claims are Metrum AI's and have not been independently verified by AMD.	FACT	B2; V8 46
R19	The quality floor is a prediction, not a guarantee.	FACT	S3; V8 51
R20	Boundary statements: a predicted floor is not a guarantee · regex PII filtering is not a guarantee · capture is off by default and export is admin-authorized and audited · the customer clears consent, retention, base-model licence and provider training terms.	FACT	S15; V8 51
R21	Frontier and open model names are as published in September–October 2026 (Claude Opus 5, GPT-6 Astra, K2 Horizon 7B, Muse Glimmer 30B, DeepSeek V4.1 Flash, Gemma 4 E4B, Qwen3.6-35B-A3B,	FACT	S9, S10, S12; V8 24, 49

	Qwen3.8 27B).		
R22	Community edition is Apache-2.0 with no licence key; the Enterprise edition adds production validation, named support and signed releases.	FACT	S20; V8 37
S-M1	Starter portfolio on one XE7745, published checkpoint footprints, not measured VRAM, and not a validated partition: K2 Horizon 7B ~18 GB (TP=1, ~126 GB gross remainder) · Muse Glimmer 30B ~60 GB · Qwen3.8 27B ~56 GB x2 (TP=1 replicas) · DeepSeek V4.1 Flash ~510 GB FP8/FP4 (TP=4, ~66 GB gross remainder — the tightest budget on the node).	VENDOR CLAIM	S9; V8 23, 49
R2b	d3 boundary: the live run was d3 26B in enforce mode, fail-closed verified when d3 was unreachable. It does not establish typical savings (1.4% on the seed set, 25% on one 8-turn take) or long-context latency.	MEASURED (scope)	V8 47; S13
R3b	OpenJev boundary: live run 25 Sep, 60 prompts x 2 on one RTX PRO 6000, 0 of 60 repeat flips; spend projected from a 15-call price sample. It does not establish production accuracy, Instinct performance, or commercial rights.	MEASURED (scope)	V8 47; S13
R4b	Savings-report boundary: 1,379 requests over 30 days, one developer, three list-price baselines, $58.06 vs $772.18 -> 81%–92.5% lower cost; the baseline is list price with no cached-input discount, so at Astra's cached-input rate the saving is about 67%.	MEASURED (scope)	S16; V8 47; B2

	The team-wide figure is 85%.		
R23	Neither decision model (d3, OpenJev) has been measured on Instinct yet; both ran on one NVIDIA RTX PRO 6000.	MEASURED (scope)	S13, S14; V8 47
F4b	One audited export contained 1,707 traces (1,186 train / 221 validation / 300 held-out); the uncaptured team was verified absent.	MEASURED (scope)	S13; V8 13
F10	0 of 2 candidates were promoted: the loop is proven, a better model is not yet.	MEASURED (scope)	S15; V8 13; B2
F11	LoRA training on governed, router-exported traces took about 20 minutes per run on Instinct MI325X (1,215 s; the NVIDIA run was 893 s).	MEASURED (scope)	S14; V8 14
G18	vLLM Semantic Router x AMD is an AMD-backed open project, not an AMD product (collaboration 16 Dec 2025; six-model playground Jan 2026; arXiv 2603.12646, Mar 2026: 50 ms, under 800 MB on MI300X).	EXTERNAL (cited)	V8 42, 44
G19	Read from public NVIDIA repositories on 17–18 Sep 2026: NeMo Switchyard pre-1.0 (v0.3.0, 22 Sep), standalone proxy demo/evaluation only, credentials and budgets in a host gateway; LLM Router Blueprint deprecated; Dynamo KV-aware router at the worker layer.	EXTERNAL (cited)	S19; V8 33, 34, 35
G20	Hosted routers offer many models with no hardware, but prompts and the routing decision leave the walls; gateways do keys, budgets	EXTERNAL (cite	S19; V8 34

	and load balancing well. OpenRouter can sit behind Metrum as an approved provider.	d)	
G21	The Linux Foundation's Tokenomics Foundation is standardising how AI cost and value are counted; the router's meter maps to its roadmap — a mapping, not a membership claim.	EXTERNAL (cited)	V8 30
C1	cost per accepted job = (inference + tools and sandbox + retries and failed attempts + verification + operating overhead) ÷ accepted jobs, with the latency requirement stated.	METHOD	PROMPT.md; COST-MODEL.md §1
C2	Illustrative workload (blueprint reference): 500 seats · 4 always-on agents per seat (2,000 agents) · 17 h/day · 30 days · 3% escalation to frontier · 12.6 tok/s sustained per agent · 46.3B agent output tokens/month · 12.1M jobs/month · ~3,704 output + ~74,088 input tokens per accepted job · node occupancy 68.7%.	ILLUSTRATION	COST-MODEL.md §3
C3	All-frontier: $419,451/month for the modelled workload at GPT-6 Terra list ($1/$5 per M, 20:1 agent input, 90% cached at 10%).	MODEL	model §4.1
C4	Hybrid on-premises: $33,111/month all-in = node $16,350 + 32 Halo boxes $4,475 + cloud burst $0 + 3% escalation $12,287. 92.1% below all-frontier.	MODEL	model §4.1
C5	All-cloud (elastic): $34,644/month (elastic MI325X GPUs at $3.80/GPU-hr + hosted open models at $0.95/M + escalation).	MODEL	model §4.1

	4.4% above hybrid.		
C6	Break-even occupancy vs elastic cloud GPUs: 74.7% (~75%) — the node's all-in monthly cost equals the cloud cost of the same agent-hours.	MODEL	model §4.2
C7	Payback vs all-frontier: 1.6 months.	MODEL	model §4.2
C8	Payback vs all-cloud: 33.1 months.	MODEL	model §4.2
C9	On-prem cost per M output tokens at the modelled 68.7% occupancy: $0.055/M marginal (power + ops only) and $0.364/M all-in (including capex ÷ 36).	MODEL	model §4.3
C10	On-prem $/M output tokens: at 40% busy $0.094 marginal / $0.626 all-in; at 90% busy $0.042 / $0.278.	MODEL	model §4.3
C11	Cost receipt, frontier-priced: one accepted job = $0.0359 (inference $0.0326 + tools $0.0024 + retries $0.0008).	MODEL	model §4.4
C12	Cost receipt, owned-priced at 68.7% busy: $0.0036 marginal / $0.0057 all-in per accepted job — ~9.9x below the frontier-priced job. Tool, retry and verification terms are identical across both columns.	MODEL	model §4.4
C13	Retry lever: a failure rate of 20% / 33% / 50% / 67% multiplies the cost per accepted job by 1.25x / 1.49x / 2.00x / 3.03x.	MODEL	model §4.5; A14
C14	Escalation lever: 1% / 3% / 10% / 25% of output tokens to a frontier API -> hybrid monthly $24,920 /	MODEL	model §4.5

	$33,111 / $61,780 / $123,212. The single biggest lever on the hybrid bill.		
C15	Cached-input lever: 0% / 50% / 90% / 99% cached input -> all-frontier monthly $1,168,980 / $752,575 / $419,451 / $344,498. Cached input dominates the frontier price.	MODEL	model §4.5; G7
C16	Price and cost inputs (from BLUEPRINT-REFERENCE.py): node capex $500,000 budgetary — not a quote; Halo box $3,999 x 32 boxes (1 per 16 users) -> total capex $627,968; DO MI325X $3.80/GPU-hr with no provisioning factor; GPT-6 Terra list $1/$5 per M tokens; agent input:output 20:1; 90% of agent input cached at 10% of list; electricity $0.12/kWh, PUE 1.4; node power 8 x 282 W measured + 1.5 kW host; ops $2,000/server/month + $20/Halo box; amortisation 36 months; 250 agents per MI350P, 12.6 output tok/s per agent.	MODEL	model §2
C17	92.5% lower cost vs GPT-6 Astra list for one team over 30 days in the demo (1,379 requests, $58.06 vs $772.18); at Astra's cached-input rate the saving is about 67%.	MEASURED (scope)	R4; S16; V8 47
C18	85% lower model spend vs the frontier API bill, measured on Metrum's own team (~25 AI developers, one deployment).	MEASURED (scope)	R4

3. Do-not-say list

1. Never show a figure for **our own project budget or production spend** — campaign budget, agent/model/ElevenLabs/GPU spend for making this work, or any build-bill figure, **including in the internal film** (`F3`, `F6`).

2. Never divide `$150.76` by SQLite successes, or by any success count, to produce a job price.
3. Never call the pilot NPU, Halo or MI350P validation, and never claim it establishes Router uplift.
4. Never present the 3 / 7 / 14 on-time loads as validated sustained capacity.
5. Never present short-turn MI350P throughput as agent-workload throughput.
6. Never present rented MI325X results as MI350P performance or as on-premises proof.
7. Never present the 85% internal model-spend reduction as MI350P TCO.
8. Never claim energy efficiency or tokens per joule — no NPU power was measured.
9. Never conflate the 96% NPU-only control peak with the 74% generation run.
10. Never claim a fine-tuned model beat the teacher — both run-2 candidates stayed in canary.
11. Never imply a Dell Halo SKU, a confirmed R7725 + MI350P support entry, or a Dell Helios product.
12. Never promise guaranteed quality or latency, aggregate atomic dollar budgets, automatic privacy from a classifier, or whole-agent containment from key revocation.
13. Never show a vendor claim as our own result, and never show an unverified catalog or roadmap item as fact.
14. Never put hostnames, machine names, internal tool names, instance types, regions, private links, credentials or campaign budget figures in any media.
15. Never show OctopAI, the pantheon roster or the mascot in the customer deck, the two customer films, or any external email.
16. Never upscale low-resolution footage and call it native 4K.
17. Never fabricate agent dialogue or a spontaneous incident; any fault injection shown must be labelled `TEST`.
18. Never show a number without its claim ID and label — including in captions, transcripts and appendix slides.
19. Never blur a prior result, a current capture and an illustration; each carries its own chip.
20. Never show a cost figure that is not an output of `shared/cost-model/cost_model.py`; never show two versions of the same cost figure. If the script's reference self-check fails, nothing downstream may be rendered.

4. Never in the media

Our own project budget and production spend — the campaign budget, the agent, model and voice-generation spend for producing this work, and any build-bill figure (claim IDs F3 and F6). The deck shows what the customer's workload costs, never what the chase cost us.