An agent job is a loop. Only the verifier decides.

What an agent costs is every attempt, retry, tool wait and verification it took to produce one accepted job.



ILLUSTRATION loop diagram

THE UNIT OF ACCOUNT

cost per accepted job

= (inference + tools + retries
+ verification + overhead)

divided by accepted jobs

latency stated

C1 METHOD

Four lookalikes it is not:

cost per token · utilization · finishing on time · verifier success rate

None of them is the bill.

C1/P4 METHOD · R19 FACT · shared/NARRATIVE.md §1-2 · adapted from the supplied seller deck (one virtual model for every agent) · no figure on this slide.

METRUM AI

The bill has five parts. Only one is the model.

Tools, retries, verification and overhead are the rest — and a failed attempt is not a cheap exit.

1 inference — MODEL
reasoning + cached input · C11 · $0.0326 per job

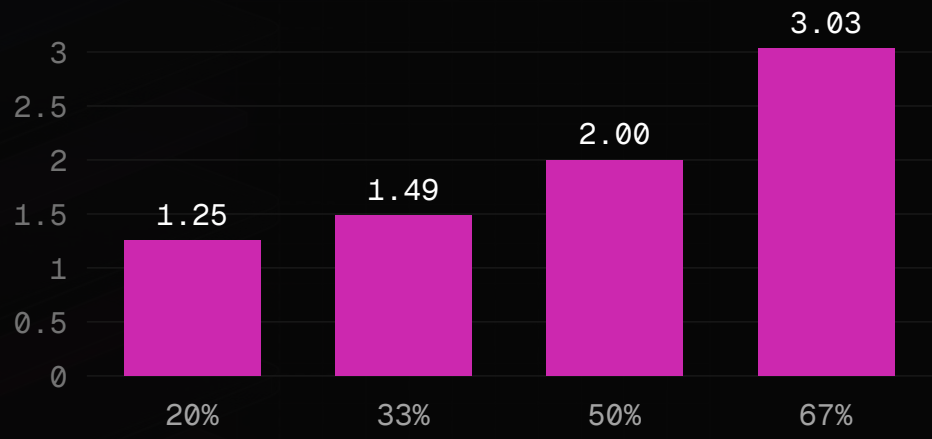
2 tool and sandbox — MEASURED
the CPU tier doing the work · A10 · 0.276 core per agent

3 retries and failures — MEASURED
failures cost real CPU · A4 · 48 vs 47 core-seconds

4 verification — MEASURED
the pass/fail decision · A14 · 4 of 5 low-load passed

5 operating overhead — NOT MEASURED
orchestrating-agent tokens · excluded, never estimated

Two tiers: 0.276 core per agent, ~250 agents per card.



A 50% failure rate doubles the cost of every accepted job.

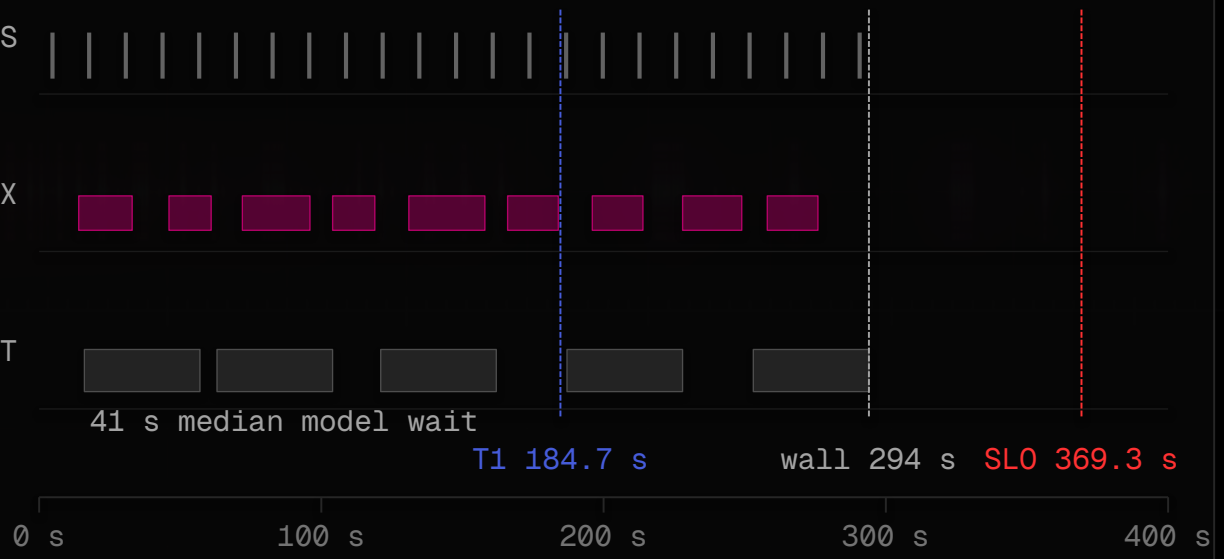
C13 MODEL on measured inputs

A4/A10/A14 MEASURED (Core Sizing Pilot, SQLite coding agents, EPYC 9R14) · C11/C13 MODEL · shared/COST-MODEL.md §4.4.

Where the time goes, and what each measure answers.

Zero throttling is not zero waiting; utilization, latency-safe load and verified success are three different answers.

ONE INSPECTABLE TASK — T1 184.7 s, SLO 369.3 s · ILLUSTRATION SHAPE



Zero of ~4,000 calls throttled — yet every run waits on the model (A5).

THREE MEASURES, THREE ANSWERS

81.9%

busy at 5 agents on 1 core

3 / 7 / 14

agents at the latency target

4 of 5

low-load trials passed

A8 · A11 · A14 MEASURED (pilot)

V1 FAIL | V4 PASS | V6 FAIL

A5/A7/A8/A11/A14 MEASURED (Core Sizing Pilot; single finite waves, not sustained capacity) · A12 verdicts registered · update v8 16/18/19.

Policy runs per agent key, in front of the runtime.

Eligibility first, prediction second, promotion is separate — the router is a gateway, not the serving runtime.

ADMIT
authenticate the key, check ceilings

FILTER
a sensitive verdict narrows the allow list

ROUTE
eligibility first, prediction second

WHAT POLICY CHANGES, MEASURED

204 / 204

prompts to the intended tier

R2 MEASURED

85%

lower model spend

R4 MEASURED

104 ms p50 and fail-closed. 204/204 does not establish typical savings, and neither decision model has been measured on Instinct yet. The 85% is our own team, one deployment — not MI350P TCO.

100% of routed tokens land on a usage row; the pod is CPU-only (R14, R15).

GUARDRAIL — NOT A GUARANTEE

The Router is a gateway, not the serving runtime.

No guaranteed quality, no atomic dollar budgets, no automatic privacy, no containment.

The floor is a prediction.

R-G guardrail

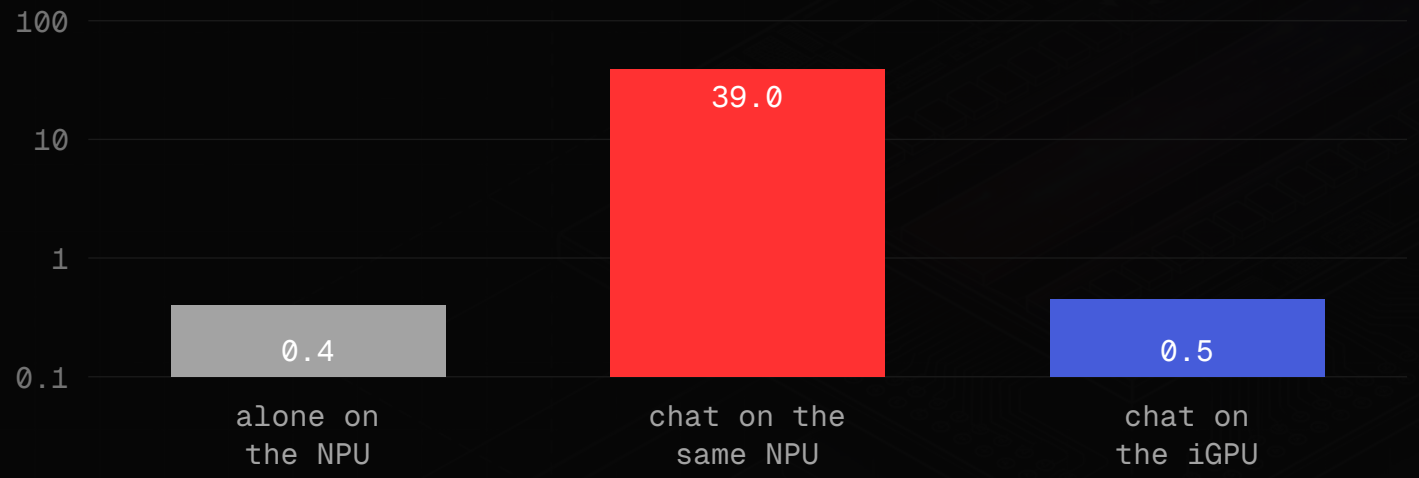
204/204 is one run, not typical savings. Neither decision model is measured on Instinct yet. OpenJev weights are CC BY-NC 4.0.

N2a-c MEASURED | N4 MEASURED

The NPU serves one request at a time.

Excellent at bounded local work — contention is a design input, not a footnote.

THE SAME DECISION, THREE WAYS — LOG SCALE, 0.1 s TO 100 s



Same run, same suite: 0.40 s alone · 39 s while chat shares the NPU · 0.45 s with chat moved to the iGPU (chat 9% slower).

N2a / N2b / N2c MEASURED (preliminary)

THE BOUNDED-LOCAL-WORK ENVELOPE

quality suite	58 / 60 vs 55 / 60
NPU busy	93% busy
Whisper	2.4 s vs 4.2 s

No power was measured: no energy-efficiency claim, ever.

The 96% control peak is a different run (N4).

FastFlowLM NPU kernels: free for commercial use; redistribution in unmodified form under its binary license (N6).

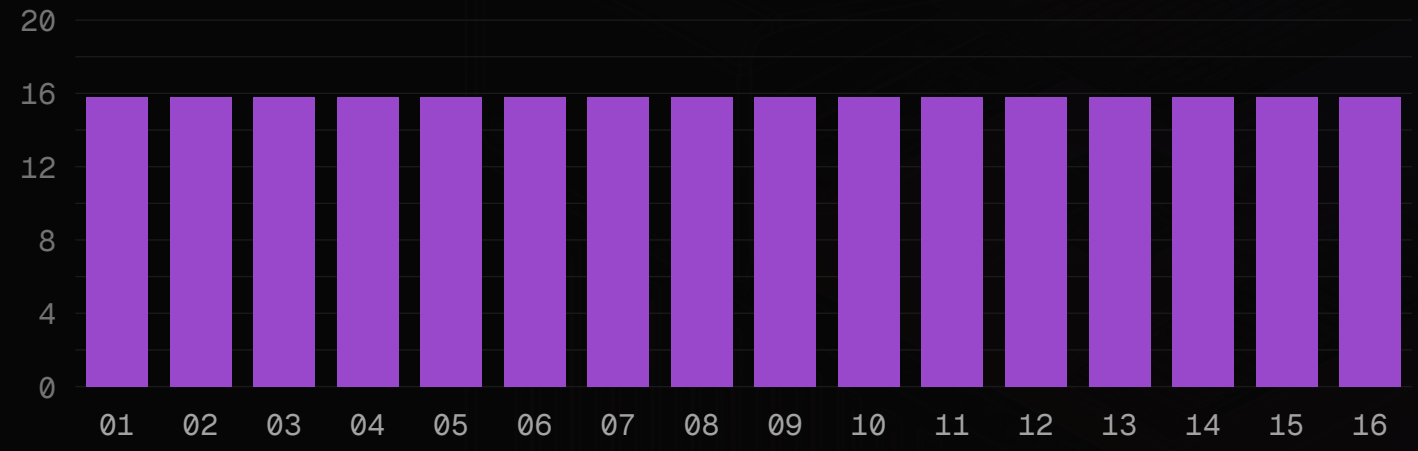
N1/N2a/N2b/N2c/N4/N5 MEASURED (preliminary, one device) · inputs/dell-amd-npu-report-preview.pdf · no source-deck ancestor.

METRUM AI

Strix Halo: a measured team-tier envelope.

16 concurrent chat streams at 15.8 tok/s each — quality level with GB10 within noise.

16 CONCURRENT CHAT STREAMS — GEMMA 4 E4B, ROCM/VLLM



15.8 tok/s per lane · 253 tok/s total · 0.86 s first token, on the official rocm/vllm build (H1).

H1 MEASURED

THE ENVELOPE AND ITS GUARDRAILS

Ryzen AI Max+ 395, 128 GB.

Not a Dell SKU — no Dell Halo product exists today. List price $3,999 (H3).

single user	53 tok/s
vs GB10	level at 16 users
users per box	16

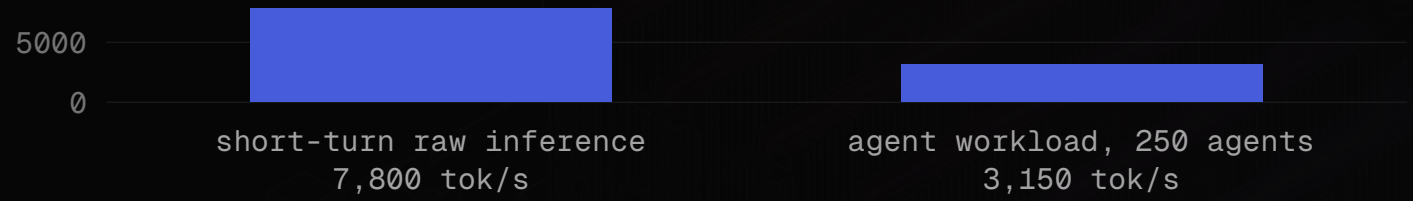
Ops time decides the economics: a few hours a month loses to a hosted API.

MI350P on Dell PowerEdge: the accepted-work tier.

~250 concurrent coding agents per card — short-turn raw throughput is not agent throughput.

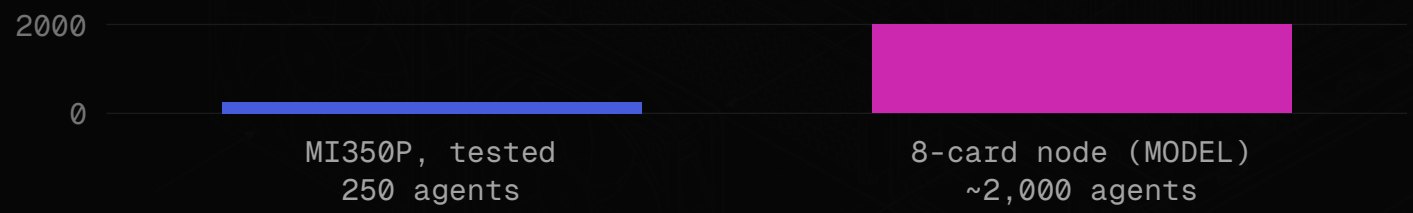
THE SAME UNIT, TWO DIFFERENT MEANINGS

I1 MEASURED (short-turn) · I2b MEASURED



CAPACITY BASIS — TESTED, THEN EXTRAPOLATED

I2 MEASURED CLAIM · I2c MODEL (extrapolated)



THE PLATFORM, AS PUBLISHED

XE7745	8 x MI350P, 1.15 TB
MI350P	144 GB HBM3E
measured draw	282 W

Dell names XE7745 and R7725 with MI350P. The catalog listing is in discussion, not live.

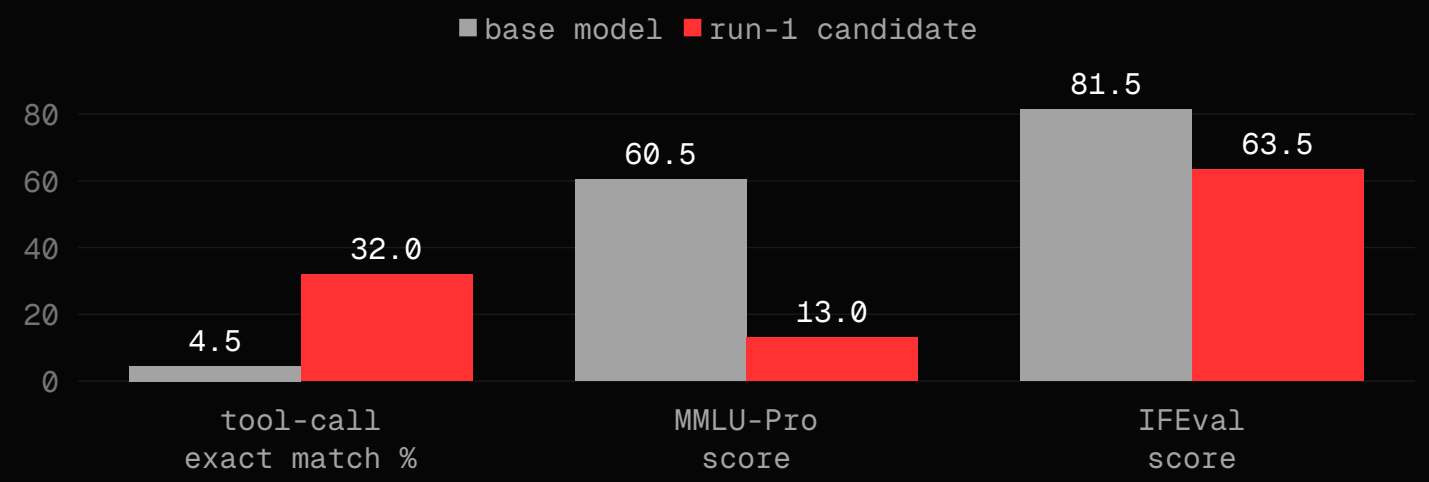
Rented MI325X is a stand-in, not on-prem proof (I3).

F1 MEASURED | F7 MEASURED | F2 MEASURED

The gate refused a candidate that looked like a win.

A 7x tool-call gain that collapsed reasoning: 0 of 2 promoted, and that is the loop working.

RUN 1 — BASE MODEL VS THE TRAINED CANDIDATE



The 7x tool-call gain is real; MMLU-Pro fell 60.5 to 13.0 (F1).

0 of 2 promoted — the loop is proven, a better model is not yet. (F10)

RUN 2 — THE GATE THAT WAS NOT PASSED

NVIDIA	+0.5 pp
AMD	-1.0 pp
required	+10 pp
outcome	0 of 2 promoted

NVIDIA 5/200 vs 4/200 · AMD 7/200 vs 9/200 (F7).

One audited export of 1,707 traces, uncaptured team verified absent (F4b).

MI325X was a rented stand-in (I3).

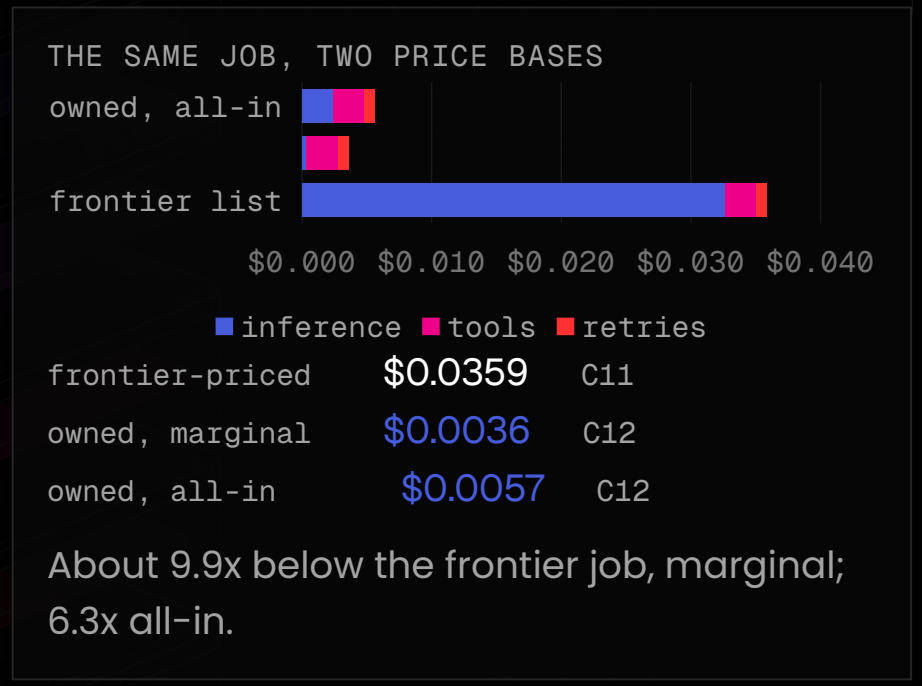
The receipt: one accepted agent job.

Token counts and the dollar column are modelled; the CPU work and the verifier outcome are measured.

TERM	QUANTITY / BASIS	CLAIM · LABEL
Inference - output	3,704 tokens per job	C2 · MODEL
Inference - input	74,088 in, 90% cached	C2/C16 · MODEL
Tool / CPU	47 core-seconds, median	A4 · MEASURED
Retries	+48 core-seconds x 25%	A4/C13 · MEAS+MODEL
Verification	verifier pass / fail	A14 · MEASURED
Operating overhead	excluded, not estimated	C1 · METHOD

divided by accepted jobs · latency stated (C1)

$150.76 whole pilot — a study cost, never a per-job price (A6).



About 9.9x below the frontier job, marginal; 6.3x all-in.

Assumptions: GPT-6 Terra $1/$5 · 20:1 input:output · 90% cached at 10% of list · core-hour $0.1875 · node capex $500,000 budgetary (C16).

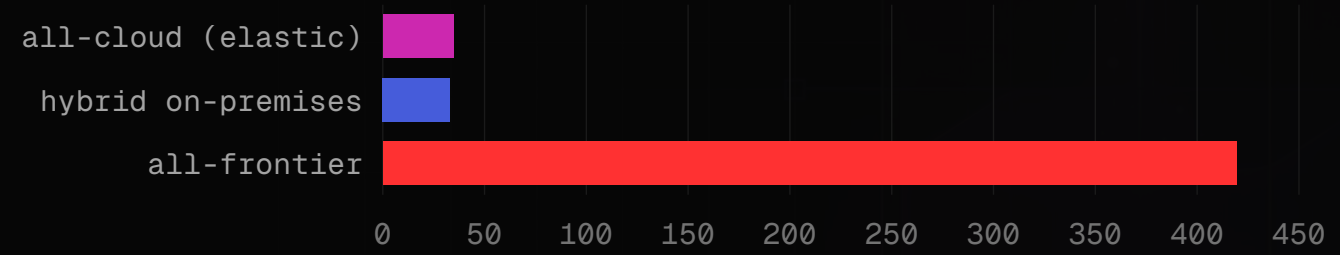
C11/C12 MODEL (shared/cost-model/outputs.csv) · A4/A5/A14 MEASURED (pilot) · C16 price inputs · shared/COST-MODEL.md §4.4.

C3-C5 MODEL C6 MODEL C16 MODEL

On-prem, cloud or frontier — with the assumptions.

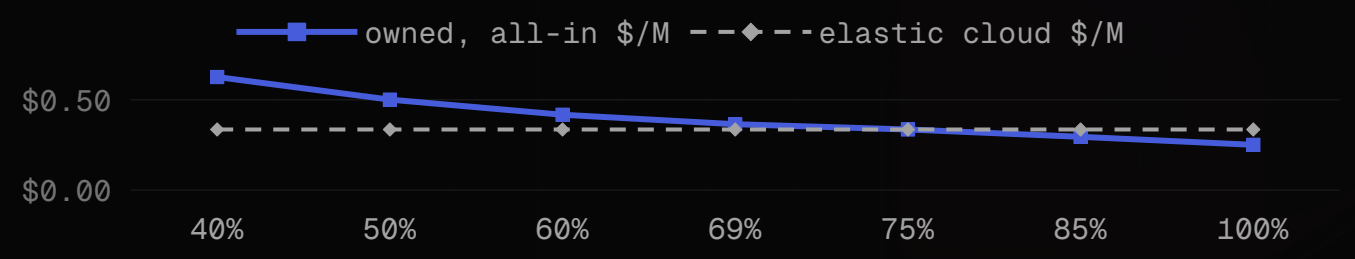
Modelled workload at 68.7% occupancy: hybrid is 92.1% below the frontier bill, and only 4.4% below all-cloud.

MONTHLY, ALL-IN — OCCUPANCY DECIDES WHICH WINS (C6)



$419,451 vs $33,111 vs $34,644 a month, all-in (C3-C5).

BREAK-EVEN 74.7% — ABOVE IT OWNED WINS, BELOW IT RENTED



PAYBACK AND OWNED TOKEN PRICE

1.6

payback vs frontier

C7 MODEL

33.1

payback vs cloud

C8 MODEL

capex $627,968 · on-prem $0.055/M marginal, $0.364/M all-in (C9).

ASSUMPTIONS THAT PRINT WITH EVERY FIGURE

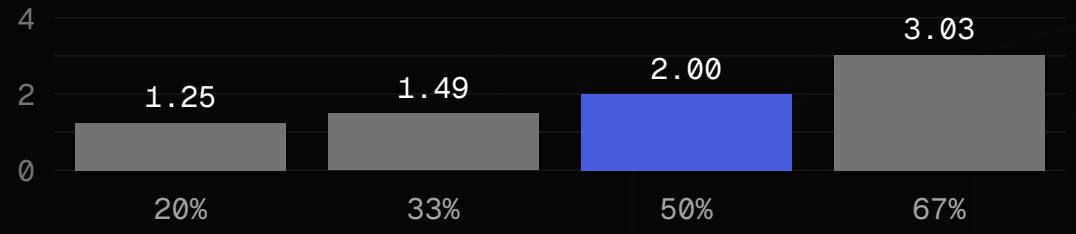
Capex $500,000 budgetary + 32 Halo boxes $3,999 = $627,968 · MI325X $3.80/GPU-hr · GPT-6 Terra $1/$5 (C16).

C2-C10/C16 MODEL, assumptions on screen (shared/COST-MODEL.md §3-§4.2) · I3 rented-stand-in caveat · blueprint page #onprem.

Four levers decide the economics.

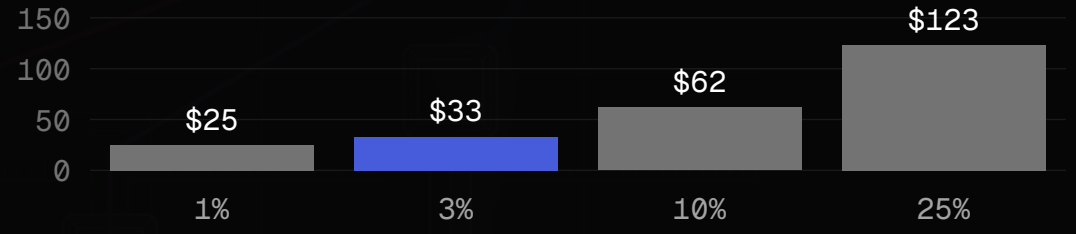
Retries, escalation, cached input and occupancy — three of them are policy, not hardware.

1 RETRIES (C13)



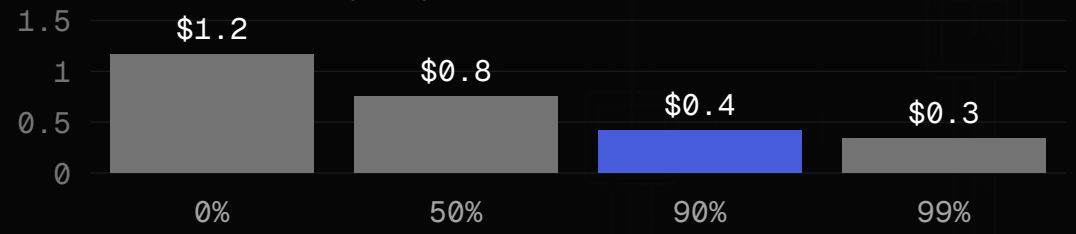
A 50% failure rate doubles the cost.

2 ESCALATION (C14)



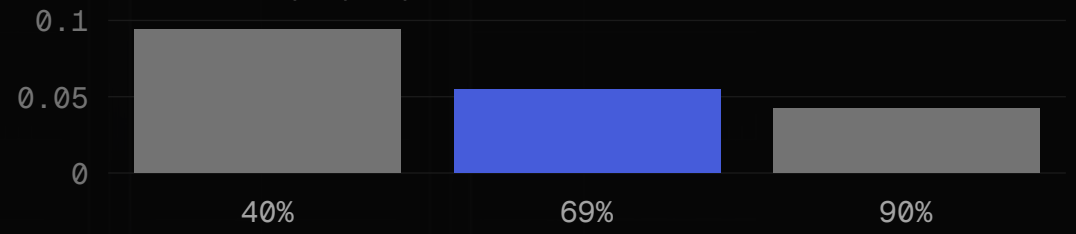
The biggest lever on the hybrid bill.

3 CACHED INPUT (C15)



Cached input dominates the frontier bill.

4 OCCUPANCY (C6/C9) — MARGINAL $ PER M OUTPUT



40% / 69% / 90% busy: $0.094 / $0.055 / $0.042 per M.

Every number re-derives. The next step is named.

Reproduction evidence, where each execution choice fits, and the proposed AMD + Dell next measurement.

REPRODUCTION EVIDENCE (A18)

- **Snapshot restore** — every file hash matched.
- **Offline rebuild** reproduced all 136 CSVs byte for byte.
- **Sealed prediction** re-scored as registered — it failed.
- **Teardown:** 0 instances, volumes or snapshots left.
- **One script, one CSV:** every figure here re-derives.

WHERE EACH CHOICE FITS

- **NPU** — bounded local decisions.
- **Halo** — a team box, beside the CPU tier.
- **MI350P on Dell** — concurrent agent work.
- **CPU and sandbox** — builds, tests, retries.
- **Cloud** — bursts and pilots, not the destination.

THE NEXT STEP (PLAN)

- **Constant-backlog test** to separate drain from capacity.
- **A second core type** and a measured 3-core point.
- **A Dell/AMD lab path** to MI350P validation.

CAVEATS

Testing by Metrum AI, not verified by AMD (R18). Catalog listing: in discussion, not live.

Appendix A1 — method and scope

What was measured, what was modelled, what was drawn — and the boundaries that travel with each.

PILOT SCOPE (A1) — PRINT IT WHEREVER PILOT DATA APPEARS

SQLite coding agents on pinned AMD EPYC 9R14 cores, Qwen3 Coder Next on Bedrock, 25-26 October 2026. It is NOT NPU, Halo or MI350P validation and does NOT establish Router uplift.

WAVES, VALIDITY AND THE REGISTERED VERDICTS (A3, A12)

The 3 / 7 / 14 on-time loads were single waves, not sustained capacity. All six sustained studies were invalid; the held-out 4-core prediction failed (V1/V6 FAIL, V4 PASS, V2/V3/V5 inconclusive).

THE FOUR LOOKALIKE METRICS (C1, P4)

Cost per token, utilization, finishing on time and verifier success rate each answer a different question. The unit of account is cost per accepted job, with the latency requirement stated. Unknown cells stay empty.

THE COST MODEL AND ITS RECONCILIATION (C16)

One script, one CSV. It reproduces the blueprint page exactly and exits non-zero if any headline drifts: occupancy 68.7%, hybrid $33,111, all-frontier $419,451, all-cloud $34,644, break-even 74.7%.

Appendix A2 — cost-model assumptions

Every modelled figure in this deck prints with these inputs; nothing is retyped by hand.

WORKLOAD (C2)

seats	500
agents	4 always-on per seat
active hours	17 h/day, 30 days
escalation	3% to a frontier API
per agent	12.6 output tok/s
output	46.3B tokens/month
jobs	12.1M per month
occupancy	68.7% of the node

PRICES AND INPUTS (C16)

node capex	$500,000 budgetary
total capex	$627,968
Halo boxes	32 x $3,999
cloud GPU	$3.80/GPU-hr
frontier list	$1 / $5 per M
agent input	20:1, 90% cached
amortisation	36 months
power and ops	8 x 282 W + 1.5 kW

CONVENTIONS AND THE CACHE NOTE

Per-M-output price bases

agent $8.80 · team chat $8.00 · background $9.00 · routing decisions $35.00

The cache convention

90% of agent input is cache reads priced at 10% of list. Assuming 10% cached overstates the frontier bill.

Rented MI325X is a stand-in, not on-prem proof (I3).

C2/C16 MODEL + ILLUSTRATION (shared/COST-MODEL.md §2-§3, shared/cost-model/BLUEPRINT-REFERENCE.py) · values from outputs.csv.

Appendix A3 — sources, objections, open items

Where every input came from, what to say when a customer pushes back, and what is still open.

SOURCES

- **Public:** metrum.ai, the router release, Dell and AMD.
- **Internal:** pilot page, blueprint, NPU preview, fine-tuning.
- **Decks:** seller deck, battlecard, update v8.
- **Cost model:** one script, one CSV.
- **Excluded:** private links and hostnames.

OBJECTIONS

- **Owned never beats API pricing.** Break-even is 74.7% occupancy (C6).
- **Isn't AMD behind vLLM Semantic Router?** Yes — it plugs in as a signal (G18).
- **Will our data train someone else's model?** Router capture is off by default; upstream training follows each provider's terms — use on-prem or no-training providers for sensitive work (R20).

OPEN ITEMS AND RULES

- **Open:** a 3-core point, MI350P validation (A20, P9).
- **Open:** neither model is measured on Instinct yet (R23).
- **Rule:** never claim a competitor lacks a feature (see their docs).

Community edition is Apache-2.0; Enterprise adds support and signed releases (R22). Testing by Metrum AI, not verified by AMD (R18).

R18/R20/R21/R22 FACT · G18-G21 EXTERNAL (cited) · A20/P9 PLAN · shared/CLAIMS.md §7-§8 · supplied battlecard 2.