TASK #435 · QWEN3.6-35B TERNARY (MAC MINI) VS CLAUDE HAIKU 4.5

The Complexity Cliff

Three tiers of task, one local model, one cloud model, and a real dollar figure at the end. The local model didn't lose gracefully — past simple recall, it mostly didn't finish at all.

55 prompts **3** tiers **2026-09-07** **2** local runs, 700 → 4096 token budget

---

SCOREBOARD

TIER	LOCAL — MAC MINI	CLAUDE HAIKU
1 · Simple factual	17/20 (85%)	**20/20 (100%)**
2 · Multi-step reasoning	16/20 (80%)	**20/20 (100%)**
3 · Strategy — completed	5/15 (33%), at 6× budget	**15/15 (100%)**
3 · Strategy — quality of what finished	12/20 rubric pts (60%)	**51/60 rubric pts (85%)**
Cost, full 55-prompt run	**free · ~2 hrs wall-clock**	**$0.05**

TIER 1

Simple factual

Single-hop, unambiguous, one correct short answer. The baseline: if a model fails here, complexity isn't the variable that matters.

EXAMPLE PROMPT · T1_04

"The chemical formula for water is…"

Expected: **H2O** (or an equivalent written form)

LOCAL 85%

17 /20

Genuine misses — wrong or never named within budget, not truncation artefacts.

HAIKU 100%

20 /20

Clean sweep. Not the tier that decides anything.

TIER 2

Multi-step reasoning

Arithmetic word problems and small logic puzzles — one verifiable answer, reachable only by actually working through 2–4 steps.

EXAMPLE PROMPT · T2_01

"A shop sells pens for £2 each and notebooks for £5 each. Someone buys 4 pens and 3 notebooks. The total cost in pounds is…"

Expected: **23** · also tested: seating-order logic, labelled-box deduction, chickens-and-cows

LOCAL 80%

16 /20

Mostly real arithmetic slips (compound interest, age problems).

HAIKU 100%

20 /20

Worked every step shown, every time.

TIER 3

Sophisticated strategy

Open-ended B2B marketing scenarios grounded in real practitioner frameworks — no single correct string, graded against a 4-point rubric per prompt instead.

EXAMPLE PROMPT · T3_01

"A B2B SaaS company has just closed a Series B and is moving from founder-led sales to a repeatable marketing motion, with a first marketing hire to make. Propose which role to hire first and justify the sequencing."

Graded on: names a specific role · stage-appropriate justification · what NOT to hire yet · ties to what the company must prove next

LOCAL — COMPLETED 33%

5 /15

At 6× the original budget. The rest ran out entirely mid-reasoning, or collapsed into repeated garbage.

HAIKU — COMPLETED 100%

15 /15

Every response finished, unprompted, well under budget.

QUALITY, OF THE 5 THAT FINISHED 60%

12 /20 pts

One genuinely solid answer; the rest generic, defaulting to textbook SEO advice on-brand questions.

QUALITY, ALL 15 85%

51 /60 pts

Specific, concrete, correctly hedged — same generic-SEO blind spot on the two AI-visibility prompts.

REAL COST, NOT AN ESTIMATE

What six hours would actually cost

The Anthropic API returns the real token count for every call. Summed and priced at Haiku's published rate, the entire 55-prompt run cost **five and a half cents**.

MAC MINI, BOTH RUNS

~2 hrs

wall-clock, £0 marginal cost — and still only 5 of 15 strategy answers finished

HAIKU, FULL 55-PROMPT RUN

$0.05

1,757 in / 13,341 out tokens · every prompt completed

Scaled to the self-refine loop's actual pace (201 cycles over 6.5 hours, roughly 3 calls per cycle) at Haiku's measured tier-3 rate: **around $1.50–$2.25** for the cloud equivalent of a full overnight local run — and unlike the local run, every one of those calls would finish.

VERDICT

Past simple factual lookup, the local model can't be trusted to finish a response, let alone a good one — even at six times the token budget, with the known collapse-prevention fix applied.

Cloud completes every time, scores higher, and the six-hour-equivalent workload costs about two dollars. The only remaining local advantage is privacy in principle, not anything this measurement showed.