01 / THE COMPANY

Private AI.
Worth owning.

StudioC makes capable AI practical on hardware a business controls — through smaller models, faster local infrastructure and a private workspace.

RECOVERED MODEL

27B

8.48 GB model file · Established build

BENCHMARK RETENTION GATES

5/5

Above 80% of teacher scores · Matched tests

CORRECTED HARNESS TOOL CALLS

300/300

287/300 before correction · Unchanged weights

Established model · GSM8K 91.9% · MMLU 97.1% · HellaSwag 100.6% · ARC Challenge 92.0% · HumanEval 81.4%

The model. The engine. The workspace. One private AI stack.

01 / COMPRESS

Anneal

Compress existing models. Recover capability through training. Compare against the original.

02 / RUN

rMLX

Our Rust and Metal runtime. The graph, kernels and serving path built for local hardware.

03 / WORK

The Harness

A private workspace for company knowledge, cited answers and tools that do useful work.

04 / ADVANCE

E-RLM

Our next research frontier: execution inside active reasoning. Protocol working; capability gains still to prove.

Search can be a business of its own.

Companies such as Exa and Tavily have built businesses around search APIs for AI. StudioC Search puts search and page retrieval inside our stack, without requiring a key to an external search API. A standalone StudioC Search API is a future commercial opportunity.

One stack. Several routes to revenue.

Local installations and support create the first customer relationship. Dedicated private hosting is planned. Search APIs and specialist knowledge subscriptions could create additional recurring revenue. These future offers would sit alongside local installs: no change to their operation and no move of customer data.

Founder-funded by Harry, Matt and Riley. Pre-seed funding advances business pilots, runtime integration and research. Harness currently uses stock models. Future standalone services would need product validation and source-rights review.

02 / ANNEAL + rMLX

Smaller weights. *Serious capability.*

A 27B model compressed to 8.48 GB, retaining more than 80% of teacher performance across five matched benchmarks. The established recovered build passes all five retention gates.

FIVE MATCHED BENCHMARKS / TWO DISTINCT RECOVERED BUILDS

Benchmark	Original correct	Established · 12 Sep	Newer · 13 Sep
GSM8K · maths	246/250	226/250 · 91.9%	222/250 · 90.2%
MMLU · knowledge	479/570	465/570 · 97.1%	461/570 · 96.2%
HellaSwag · commonsense	685/1000	689/1000 · 100.6%	693/1000 · 101.2%
ARC Challenge · science	363/500	334/500 · 92.0%	334/500 · 92.0%
HumanEval · code	129/164	105/164 · 81.4%	94/164 · 72.9%

Each cell shows correct answers and retained teacher score. Established build: 5/5 gates above 80%; newer variant: 4/5, with coding recovery in progress. GSM8K: 8-shot, strict match, 1,024-token cap. MMLU: 5-shot. HellaSwag: 10-shot; ARC: 25-shot, normalized accuracy. HumanEval: unfiltered pass@1, 1,024-token cap. Matched subsets, materialized BF16 on CUDA; separate from native execution.

NATIVE EXECUTION / M1 MAX · METAL

16.58 tok/s

At 4K context · 64 generated tokens
Newer recovered variant · One measured run

125.26 tok/s prompt processing · 24.07 s load.

8.48 GB model file · 10.707 GiB resident weights · 11.936 GiB peak RSS. Memory and long-context optimisation remain in development; sub-10 GB total runtime is not yet achieved.

Metal numerical check: mean forward KL 0.12324, argmax agreement 84.18%, 1,024 positions across four contexts. Passed its comparison gate; separate from task accuracy.

COMPONENTS VS MLX 0.32.2 / M1 MAX

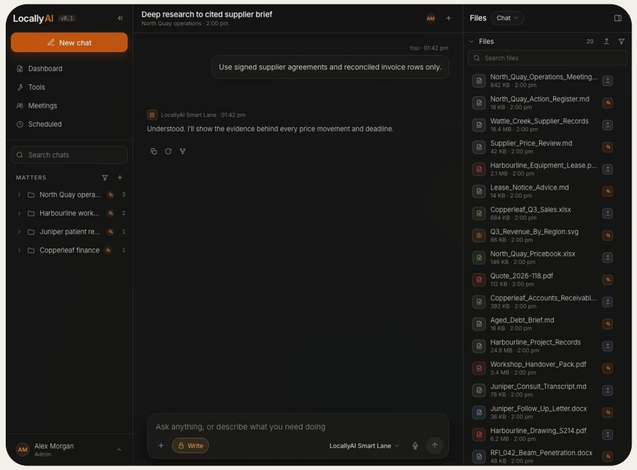
Paired workload	Speed ratio
Graph building	3.07×
Dispatch overhead	1.90×
Fused 2,000-operation chain	7.95×
64 MB tanh-GELU activation	6.48×
fp16 4096³ matrix multiplication	1.00×
Decode microbenchmark	0.93×

Paired medians, n=14; 174/174 parity cases. Component gains are separate from full-model speed. The decode microbenchmark is 408.5 vs 437 tok/s.

03 / THE HARNESS

From a question. *To work done.*

Private AI that can actually do work. Company knowledge, tools and approvals in a workspace deployed on hardware the business controls.



Actual Harness footage · Fictional demonstration records

Find. Make. Follow through.

Cited answers and source passages. Reusable inline tools and spreadsheet charts. Meeting transcripts, owned actions and scheduled reports on your machine.

StudioC Search, Mail & Connect.

Web search without an API key to an external search provider. Direct mail integration with Outlook, Gmail and IMAP hosts. Certificate-based remote access to your appliance.

Industry context. Company control.

Versioned source references, workflows and templates alongside company files. Read-only, approved writes or broader access per chat. External connections stay optional.

Knowledge worth building on.

401 source references

85 workflow guides

60 templates

Federal + all state/territory settings: NSW / VIC / QLD / WA / SA / TAS / ACT / NT

Australian national core
Privacy, consumer law, employment records and shared business obligations.

Healthcare & NDIS
Privacy, consent, clinical records and provider obligations.

Accounting & tax
ATO guidance, GST/BAS, payroll, super and practice records.

Construction & trades
Contracts, licensing, WHS, variations and payment claims, with NSW guidance.

NSW legal & conveyancing
Conveyancing law, client identity, costs disclosure and settlement checks.

NSW real estate
Agency and tenancy rules, disclosure, trust money and property management.

287 → 300/300

Automatic correction of malformed structured tool calls — without changing model weights. Same 300 prompts, same stock 4-bit model.

12/12

Exercism-based code set. Tests rerun on a clean checkout. A separate, small internal test.

Start with the workflow.
Closed pilots in preparation. LocallyAI AU handles the appliance, on-site setup, accounts, groups and support. Trained contractors are the planned route to wider delivery.

A clear release path.
Code availability and free personal/business use are planned; licence terms remain in development. Embedding in another product will require approval and may carry a fee.

Draft v2026.08.1 · Source coverage varies by field/state; strongest in NSW. Professional review and signing pending.

Built to support professionals. Professionals in their respective fields verify current sources and make final legal, clinical, tax, financial and compliance decisions. The packs support research and drafting; they do not replace professional judgement.

04 / E-RLM + WHAT COMES NEXT

Thinking that *executes*.

Our next research frontier. E-RLM explores a different model/runtime boundary: deterministic execution inside active reasoning, with returned values carried into subsequent thinking and full files assembled at the end.

E-RLM 01 / PHYSICS

The calculation happens inside the thought.

Orbital radius	26,571 km
Motion per day	-7.21 µs
Altitude per day	+45.72 µs
Net clock drift	+38.51 µs a day

E-RLM 02 / CODE

Write. Run. Refine.

```
F8 Test identifies index + stopping faults
F9 Corrected result: (4, None)
KEEP F1 · F2 · F5 · F6 · F7 · F9
FILE utils.py assembled by the runtime
```

The original two scripted demonstrations. Orange values and final files come from the runtime, never re-typed by the model.

RESEARCH FRONTIER / CAPABILITY PROOF PENDING

A working protocol. The next proof ahead.

The execution protocol and runtime are working. The current fine-tune has learned the execution notation, but has not yet demonstrated improved problem-solving or reliable execution selection.

Next proof Held-out problems / execution selection / correct answers / clean stopping / generalisation.

A foundation we can test.

The Rust controller operates inside decoding; an isolated Python/SymPy worker executes operations. Both hosts passed 61 runtime gates and eight focused recovery tests. The original demonstrations show the intended interaction, not a reasoning benchmark.

Fund the platform. Advance the frontier.

Business pilots and compressed-runtime integration lead the next product milestones. E-RLM follows an explicit capability gate: demonstrate useful, reliable execution on unseen problems before making a product-performance claim.

13 September 2026. Training completed: 141 steps, six epochs, 110 corpus replays; validation NLL 2.257 → 0.682. These are process measurements. Initial matched pilot: maths 7/8 versus base 8/8; reasoning 2/4 versus 4/4; neither selected execution. Small sets, full campaign incomplete. Earlier validation contamination and conflicting training targets prompted data repairs; the measured adapter has not established a reasoning gain.