

A 27B model. *In 9.03 GB.*

StudioC makes capable AI practical on hardware a business controls.
Smaller models, a native runtime and a private workspace, built together.

CAPABILITY RETAINED

97.35%

Average across seven core benchmarks.
Original full-precision model = 100%.

Approximately 96.00% across the broader twelve-task set

Same model. Almost six times smaller.



Model files in decimal GB. Package size is separate from running memory.

Less hardware between a business and useful AI.

Our starting point is private business AI: company knowledge, tools and approvals on a machine the customer controls. The ambition is infrastructure that makes capable AI practical far beyond the cloud.

Model-only thinking-on results. Peak memory for model + 32K context: 11.69 GB on Apple silicon, reproduced in 3 of 3 runs. Acceptance on a real 16 GB machine remains open.

About 3.3 A100 GPU-hours (US\$5.30) for data capture, compression and recovery. Setup and data generation excluded. Full benchmark run: US\$2.27.

Harry · Matt · Riley

FOUNDERS / HUNTER REGION, AUSTRALIA

Better on the web.

Watch the Harness videos and animated E-RLM demos.

studioc.ai/pitchdeck



Smaller weights. *Measured capability.*

Longer thinking-on recovery brings the same 9.03 GB package closer to the original model. Native execution, memory and reliability are checked separately.

Recovery brings capability back.



Selection set: down 63%. Independent confirm set: 0.625 to 0.254, down 59%. Lower is closer to the original model.

CURRENT MODEL · SEVEN CORE BENCHMARKS

Benchmark	Retention
MBPP	98.91%
HumanEval	96.10%
GSM8K	98.98%
MATH-500	93.98%
ARC-Challenge	98.64%
MMLU	96.18%
IFEval	98.68%

Model only, thinking on. Core mean 97.35%; lowest 93.98%. Broader twelve-task mean ~96.00%. Retention averages per-task score ratios, not answer accuracy.

Recovery: scales + norm weights. Same trainable set, previously 0.71%; not recalculated. Codebook and indices fixed; 2.69 bits/weight.

M5 MAX GENERATION

95.1 tok/s

One sampled 27B server run: 4,361 tokens, MTP3 and correction on. Separate code task: 92.1 tok/s. Single runs; decode only.

32K ATTENTION + CACHE

38.96 to 2.165 ms

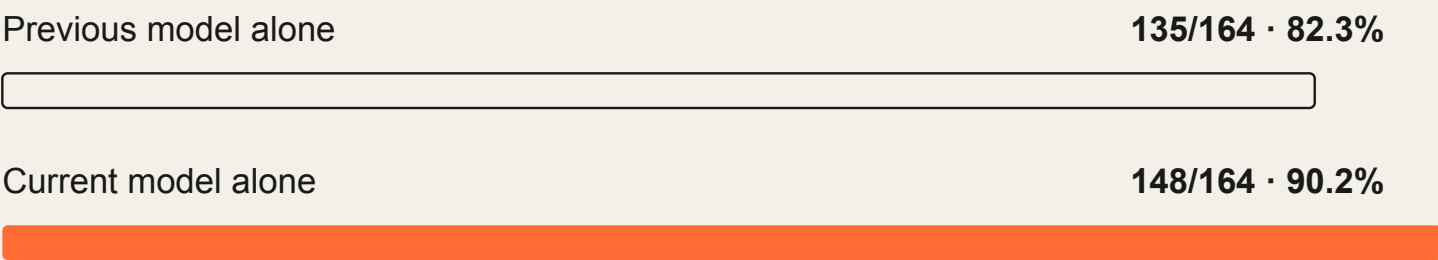
About 18× for one operation, not the model. 4-bit cache: about 3× less space than 16-bit, at similar kernel speed.

NATIVE CHECKS

19/20 · 7/10 · 44/60

32K retrieval / constraints / tool tasks. Tools: original 45/60, retention 97.78%. At 8K: 20/20 retrieval. 16 GB hardware acceptance open.

A stronger model before correction.



HumanEval answer accuracy, thinking enabled. Neither model uses the correction policy in this comparison.

CURRENT MODEL ALONE · CORRECT ANSWERS

Task	Current	Original	Retention
MBPP	91/100	92/100	98.91%
MATH-500	125/150	133/150	93.98%
HumanEval	148/164	154/164	96.10%
GSM8K	97/100	98/100	98.98%

Correct answers / problems, compared with the full-precision original. These are current-model scores without correction. Overall corrected retention is not yet available for this model.

New-model output-limit cases: 13/164 HumanEval and 14/150 MATH-500, versus 27/164 and 26/150 previously. HumanEval repeated blocks: 3/164 versus the original model's 1/164. The repetition check did not pass.

Divergence now uses thinking-on examples up to ~40K tokens, vs 1,024 before. Compare each test's reduction, not raw values across tests.

Better on the web.

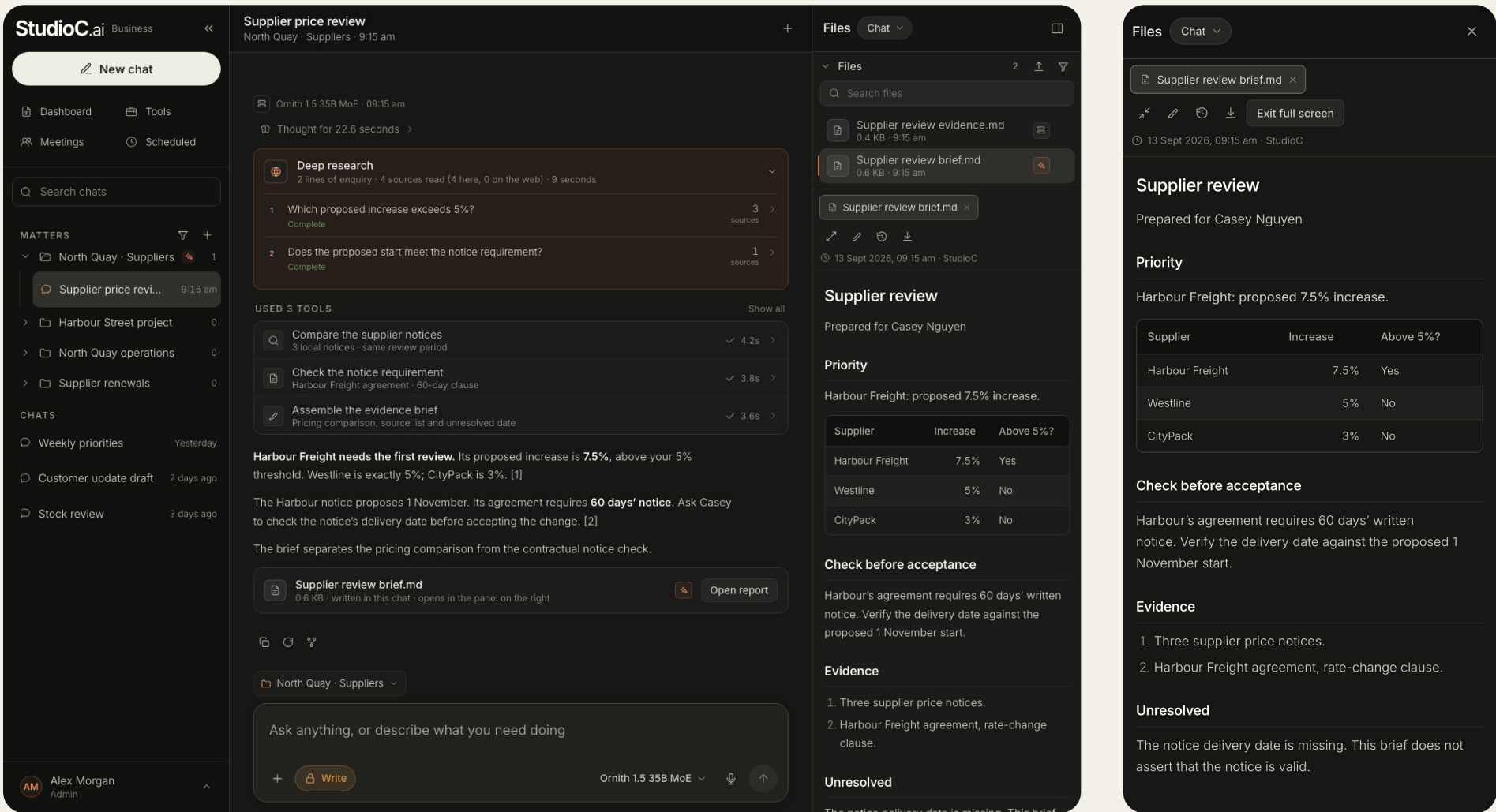
Watch the Harness videos and animated E-RLM demos.

studioc.ai/pitchdeck



Private AI. *Real work.*

For teams whose work lives in contracts, spreadsheets, email and company records. The Harness puts their knowledge, conversations and tools in one private workspace.



4:3 DESKTOP + PHONE / WATCH BOTH STYLES ON THE WEB

Fictional records and pre-generated responses. Videos are sped up and edited for demonstration.

Six Australian knowledge packs

National core · Accounting & tax · Construction & trades
Healthcare & NDIS · Legal & conveyancing · Real estate & property

Packs support research and drafting. Qualified professionals verify current sources and make final decisions.

Find it. Check it.

Cited answers and source passages. Compare agreements, analyse spreadsheets and turn meetings into owned actions.

Useful tools. Human control.

Reusable inline tools, scheduled reports and reviewed writes. Permission checks live outside the prompt.

Local, with doors.

StudioC Search needs no external search-provider API key. Mail connects directly to Outlook, Gmail and IMAP. Connect gives staff certificate-based access to the appliance.

401 sources · 85 guides · 60 templates

Federal + every state and territory selectable. Development inventory; coverage varies and is currently NSW-heavy. Professional review and release signing pending.

Better on the web.

Watch the Harness videos and animated E-RLM demos.

studioc.ai/pitchdeck



Start local. *Build much further.*

Bring private AI into real businesses, then build on the infrastructure underneath it. A commercial path grounded in useful work, with room for a much larger company.

THE FIRST COMMERCIAL ROUTE

Machines, software, local experts.

LocallyAI AU plans appliance supply, installation and ongoing support across Australia through local experts. Start with a team's actual workflow, a controlled pilot and clear acceptance checks.

INFRASTRUCTURE BEYOND THE APPLIANCE

Search is a business in its own right.

Exa and Tavily built businesses around search APIs for AI. StudioC Search brings search and page retrieval into our stack without an external search-provider API key. A standalone API, private hosting and professionally reviewed knowledge subscriptions are future opportunities alongside local installs.

SERVER-INTEGRATED RELIABILITY

12 components · 65/65 tests

1,876 replayed generations · 0 content violations.
Correction on: 4/4 control outputs stayed identical.

First live runaway corrected.

8,192 tokens of unfinished thinking led to a right answer after correction. One observed success; the larger runaway test is running.

E-RLM / THE NEXT RESEARCH FRONTIER

Thinking that executes.

```
01 / PHYSICS  
  
Motion -7.21 · Altitude +45.72 µs / day  
  
Net drift +38.51 µs / day · Carried forward.
```

```
02 / CODE  
  
F9 Corrected result: (4, None)  
  
KEEP F1 · F2 · F5 · F6 · F7 · F9  
FILE utils.py assembled by the runtime
```

Two scripted demonstrations. Orange values and final files come from the runtime, never re-typed by the model. The execution protocol works; a general reasoning gain and reliable execution selection remain to be demonstrated on unseen problems.

What we prove next.

16 GB at 32K. Full-length reliability. Secure Harness pilots. Broader platform support. Regular models are supported; compressed-model installations await acceptance.

Better on the web.
Watch the Harness videos and animated E-RLM demos.

studioc.ai/pitchdeck