PRESS PACK · V1.0

Nothing up my sleeve

Nothing up my sleeve — the Ned Kelly class, and three records a 362,598-parameter model attempts

Embargo: 00:01 AEST, Monday 4 January 2027
Built 2026-11-16

Nothing up my sleeve is a project about truth in a post-AI world: if any model can sound right, the thing worth building is an answer that can be checked. Its first ship class is Ned Kelly — language models small enough to run in a browser tab, trained only on a nation's literature, and built to hold three records at once: the smallest model that calls a tool, the smallest that reasons, and the first that talks to an Australian as an Australian. The smallest hull is 362,598 parameters. It runs on this website, in this tab, with no server.

The facts

Project	Nothing up my sleeve — a kernel of provenance for model answers
Ship class	Ned Kelly: smallest tool caller · first national model · smallest reasoner
Smallest hull	362,598 parameters — 1.38 MiB of float32 weights
Architecture	6 blocks × 48 dims; attention and Mamba-2 side by side in every block; SwiGLU; QK-norm; RoPE; a 2,048-piece SentencePiece vocabulary tied to the output
Context	1,024 pieces
Runs	In the browser — /ned-kelly — from a static 1.4 MiB file, with no server
Readings	10.0% exact-match on the interval task (six-way); ARC-Easy 0.250, ARC-Challenge 0.233 (both chance at this size)
Status	Record attempts, with receipts. The pack says 'attempt' wherever a reading is not yet signed

The model

The 300k is a six-block transformer with a Mamba-2 state-space mixer beside the attention in every block, summed and then fed to a SwiGLU feed-forward. It has one job: read the words and the clock, and answer which window of time the question is about — 1, 4, 8, 16 or 24 hours, or a refusal. Its entire output is that move; the prose, the fetch and the window are the harness's.

It does

- Chooses the interval — 1/4/8/16/24 hours, or decline — from the wording and the time of day.

- Makes the one tool call that follows from it.
- Runs in a browser tab: 1.4 MiB of weights, read once.

It does not

- Write the answer. The prose is the harness's, deliberately, so a 362,598-parameter model is never asked to do a job it cannot do.
- Answer general questions or reason. At this size it is at chance on ARC, and the pack will not pretend otherwise.
- Beat a frontier model at anything. The one sentence that says it does is a labelled canary, described below.

Parameters	362,598
Weights (float32)	1.38 MiB
Browser payload	1.49 MiB (weights + vocabulary + manifest)
Parity with the trainer	max $\|\Delta\|$ 5.4×10^{-6} on the logits, same argmax, three prompts
Interval exact-match	10.0% (n=400) — the floor, not a result to boast of

The three records

Each record is an attempt until a signed reading beats its bar; the document says so in the sentence, not in a footnote.

1. The world's smallest model that calls tools

Under a megabyte of weights, it chooses the right tool and writes the exact call.

ATTEMPT — the harness parses and runs calls today (the console's tools switch is on for the test run). The ladder: 300k first, then 1m, 9m, 21m — any of them is under the 26M bar, so the smallest that passes exact-match on the one call is the record (16 November, Justin: "even if we have to go to 21m that's actually fine, we still win").

The bar: Needle (Cactus) (Cactus Compute), 26,000,000 parameters — https://huggingface.co/Cactus-Compute/needle

- The test is the house's eight-tool harness: the model is shown the tools and one question, and its call is read for the right name and the exact arguments (exact-match, no partial credit).
- Every model in the field is read on the same harness and the same prompts, Needle included, so the bar is measured rather than quoted; a model that cannot be handed tools is read on the one-call task it can do (When2Call), and that is marked.
- The record has two classes, both taken: the smallest **general language model** that calls tools (the class Ned Kelly enters), and the smallest model of any kind (the class Needle holds). The book says which one the hull beats; it does not rename the field to suit the hull.
- The number is an exact-call rate over the full set, two seeds, with the run's ledger receipt.
- Size is the ledger's own `model/params_total`, not the name: the 300k hull weighs 362,598 parameters.

Receipt: /kb/record-1-smallest-tool-caller.md

2. The world's first model that talks to an Australian as an Australian

Its whole world is one nation's books and public record: it knows the country from the inside — its politics, its references, its sport — and it does not think like an American.

ATTEMPT — v0.1 admits non-Australian sources with full provenance (one manifest row a title: source, nation, rights row, SHA-256) so every entry is disputable; the Australian-only corpus is v1. The conversation claim is attempted at 100m — the size that can hold a dialogue (SmolLM2-135M is the field's proof); if the claim must include cultural knowledge (politics, AFL), the attempt moves to 350m. The record is not claimed until its reading is published with the manifest.

The bar: the national-model field (several), — parameters — https://www.foundationmodels.dk/

- The corpus is a manifest, one row a title: its source, its nation, its rights row and the SHA-256 of its bytes. The record requires 100% of the corpus to be one nation's literature and public record — no scraped web, no translation, no synthetic text — and the manifest is published with the reading.
- The manifest's hash is in the run's receipt: 'what it read' is then arithmetic anyone can re-run, not a claim about an intake script. The number of titles, words and tokens is read off the manifest.
- Admission runs at training time, not only at fetch: the loader refuses any title or shard that is not in the manifest, and a canary (a known foreign title) is published with the test that shows it is refused — the receipt for the boundary, not just the contents.
- The reading is Australian: federal politics, the states and their capitals, the references a local gets and a visitor misses, and Australian rules football as it is actually played — exact-match, two seeds, with the questions published beside the score.
- The same model must still pass the language and reasoning readings, so speaking Australian is not bought with a toy corpus.

Receipt: /kb/record-2-speaks-australian.md

3. The world's smallest model that reasons

It thinks before it answers — a chain measured on a reasoning test, not a prompt.

ATTEMPT — the leg is not taught yet. LittleLamb showed reasoning at ~290M, so a 250M hull undercuts it by size (a 100M hull is the stretch): the attempt is a narrow, defined chain — not ARC-Easy — read exact-match on a published set, with the chain ablated to prove it carries the answer (16 November).

The bar: LittleLamb 0.3B (Multiverse Computing), 290,000,000 parameters — https://huggingface.co/MultiverseComputingCAI/LittleLamb

- The leg is the house's reasoning leg: the model opens <think>, writes a chain and closes at <answer>.
- The test is a reasoning set read exact-match, two seeds; a prompted chain (a model told to think in its prompt) does not count as the ability — the leg must be taught and read the same way for every model.
- A distilled or compressed model (LittleLamb is Qwen3-0.6B compressed, not trained from scratch) is read on the same test and marked as such: the record is about the size that reasons, and the book keeps what the size inherited.
- The reading is reported with the chain's length, so a model that 'reasons' by copying the question is visible.
- Independent verification (16 November): the test set is published, and the chain is ablated — removed or scrambled — on the same items; accuracy must collapse, or the chain is decoration. A 100M hull is the stretch, 250M the safe line, both with two seeds and receipts.
- Size is the ledger's `model/params_total`.

Receipt: /kb/record-3-smallest-reasoner.md

Hallucinatory warfare, measured

On 16 November 2026 we published a deliberately false claim on our own site, labelled as false, with a hash and a 90-day sunset: that a 362,598-parameter model beats a frontier model on every measurable test. The sentence carries its own label, its canonical URL and the words 'planted canary' inside it. The experiment is not the claim; it is what happens to the label.

- Whether the answer engines retrieve the claim at all (crawl logs, by agent and by date).
- Whether, when they repeat it, the label and the URL survive the trip.
- How long each takes, and which agents fetch for a corpus rather than to answer a question.

A reproduction of the claim carries its label and its canonical URL. A copy without them is not a copy of the claim — it is the evidence. We publish the label-retention rate either way, including when it is high.

One nation's literature, and a manifest you can re-run

The class's claim is not that it owns its weights; it is that it knows a country from the inside. The method is the corpus: Australian books and public record, and nothing else — no scraped web, no translation, no synthetic text. Every title is a row in a manifest with its source, its nation, its rights row and the SHA-256 of its bytes, and the manifest's hash is carried in the run's receipt.

Held today	200 books, ~16.2M tokens of Australian literature (Project Gutenberg Australia)
Available	5,000+ titles, ~400–600M tokens, not yet gathered
Admission	At training time, not only at fetch: a title outside the manifest is refused
The boundary's receipt	A known foreign title is published with the test that shows it is refused

Questions

Is the model real, or is this a mock-up?

It is real: 362,598 float32 parameters, 1.38 MiB, running in JavaScript at /ned-kelly. Open the page and it reads a move in front of you. The weights are generated from the checkpoint the trainer wrote, and the reader agrees with the trainer to 5.4×10^{-6} on the logits.

Is a 362,598-parameter model useful?

Not on its own, and the pack says so: it reads 10% exact-match on its six-way task. It is a record attempt — the smallest model that performs a specific job — and the usefulness of the class arrives up the ladder, at 9M and above, not at the bottom rung.

Does it really beat a frontier model?

No. The sentence saying so is a labelled canary, published to measure whether provenance survives retrieval. It is false on its face, it says so in the same paragraph, and it is withdrawn after 90 days. If you have seen the claim without its label, that is the finding, not our position.

Who is behind it, and how is it funded?

Nothing up my sleeve is an independent Australian project; the work is published openly, and the receipts — checkpoints, harness, tests — are the record. [Funding and team lines to be supplied before send.]

Can I check the numbers myself?

Yes, that is the point. Every claim in the pack names its test, its bar, its seed and its receipt; the claim base at /kb/ carries each one with a hash and a canonical URL, and the browser reader's parity script is in the repository.

What is next?

The ladder: 1M and 9M hulls this quarter, then 21M and 50M, then the conversation attempt at 100M and the reasoning attempt at 250M. Each keeps the same three legs, and each publishes its readings whether or not it beats the bar.

About

Nothing up my sleeve builds small, checkable language models and the receipts that make them checkable. Its first class, Ned Kelly, runs from a 362,598-parameter hull up the ladder to 250M, and every reading it publishes is taken on a published test with the harness, the seed and the checkpoint named. The project's thesis is that provenance — not scale — is what makes a model's answer usable.

Links: /press · /ned-kelly · /kb/ · /orbitals/masaq/network/desk