

01 / How our tests work

Kira Labs | A plain-language guide for university conversations | Evidence as of October 16, 2026

A useful opening answer

"We have a tested skin-movement and restoration control, and a working numerical-test pipeline. NewBrain has made a recorded parameter update, but we have not demonstrated a new learning gain. The last independently measured benchmark is still 2 out of 12. I am seeking collaborators and computing access to test the next stages rigorously."

What has actually been shown

Area	Saved result	What it establishes
Learning benchmark	2/12: absence 2/2; facts 0/4; numerical 0/6.	The last measured score. No improvement beyond it is established.
Numerical prerequisite	A fixed NumPy 2.4.6 calculation returned 2016.	The numerical library ran and its result was captured. It does not establish learning.
One teaching diagnostic	8,837 of 11,914 finite parameter values changed. The full run failed its runtime checks.	An update path was observed. New acquisition and benchmark improvement were not qualified.
Weighted skin control	512 sampled skin vertices moved; final recorded coordinates and pose channels restored exactly.	Controlled geometry in the tested forearm patch.
Lungs, nerves and Qwen	No qualified lung/brain or skin/brain response. Qwen untested.	These remain development and test objectives.

Three labels to keep separate

Prepared means code, models or a procedure exist. **Ran** means an execution occurred, which can still fail. **Passed** means the declared checks were met in a stated scope and the saved evidence and cleanup were reviewed.

This guide explains existing saved results and future test plans. Creating this PDF did not run another scientific test.

02 / How an automatic test is built

A fixed question, a controlled procedure, saved observations and a reviewable conclusion.

1. State one question

Examples: "Does this numerical library return the expected value?" or "Does moving the forearm bone move the weighted skin and restore it?" Learning requires a different question: "Did the model acquire information it did not answer correctly before teaching?"

2. Freeze the recipe

Record the source version, inputs, starting state, expected outputs, tolerances, resource limits and pass/fail rules before running. Protected learning questions and answer keys remain with a separate scoring role. Each attempt has its own run identifier.

3. Check the starting conditions

Pause parallel experimental work and the scheduled background work. Check the current permission/gate records, an unused run slot, output location and stop condition. The reviewed NewBrain runtime uses at least 8 GiB of available RAM at startup, a 3 GiB reserve and bounded process, memory, I/O and time limits.

4. Run and record

Scripts start the declared processes, execute the selected operation and save the available raw results, errors, timestamps, state identities and resource observations. Numerical, model, skin and transport tests have different inputs and checks; one does not automatically qualify another.

5. Clean up and account for the run

Keep the background tracks paused. Complete the declared shutdown, inspect the owned process family and preserve missing or uncertain evidence as a failure or limitation. A failed attempt is retained; it is never erased or relabeled as a success.

6. Review before claiming a result

A separate project role examines the saved execution evidence against the declared rules, followed by a final evidence review. Only then are the qualified result and its scope reported. A follow-up or replication needs a fresh run record and gates.

What the resource numbers mean

Available RAM is memory free for new work, not total installed RAM. **8 MiB report limit** concerns saved evidence size, not model memory. The current reviewed NewBrain limits also retain a 40-second child / 45-second whole-run deadline and a configured numerical-provider capacity of two. Exact active numerical-worker count is not established.

Reviews described here are separate roles within the project. They are not an external university replication, peer-reviewed publication or clinical validation. New source versions need their own applicable qualification.

03 / How the learning results were obtained

A correct answer after teaching counts as new acquisition only if it was not already correct before teaching.

The last twelve-check score

The saved benchmark breakdown is **2/2 absence checks, 0/4 fact checks and 0/6 numerical checks: 2/12 total**. Absence checks assess withholding an unsupported answer. They do not establish that the model learned the four facts or solved the six numerical cases.

The exact historical benchmark state has not yet been fully joined to the newer nominated diagnostic state. Therefore a result from the current diagnostic cannot simply be relabeled as a new score on that original twelve-check benchmark.

What the latest teaching diagnostic measured

The public diagnostic asked about an answer that was already correct before teaching. One training update was saved. Review compared the model parameters, counters, objective and generated answer before and after that update.

Measurement	Before	After
Generated answer	a	a
Correct-answer probability	0.9991045212	0.9990638723
Total training objective	7.8074396472	7.5016123944
Changed parameter values	Compared with starting state	8,837 of 11,914 finite values

A parameter is an adjustable model number. The **objective** is the training formula being minimized. Here that objective improved while correct-answer probability decreased slightly. The answer was unchanged and already known. This is evidence of an update, not evidence of acquiring a new fact.

Why the complete run was rejected

The outer process review could not fully account for one process generation across an observed transition. The execution exited with failure and some final worker exports were missing. Its diagnostic observations remain preserved, but the whole run has **no runtime, acquisition, retention or benchmark qualification**. It was not recorded as another low-RAM startup stop.

The next planned learning comparison - unrun

A separate newly blinded set will contain four fact cases and two absence cases. The plan is baseline queries, four actual training updates and post-teaching queries, with private scoring and aggregate-only results. An answer already correct at baseline gets no new-acquisition credit. Fresh-process retention and generalization require further tests. This six-case study does not replace the historical twelve-check benchmark; Qwen still needs its own qualified comparison.

04 / How the skin result was measured

A controlled forearm movement, with saved coordinates at five commanded states.

The question and the control

The test asked whether a forearm bone moves its weighted skin and whether the original state can be restored. A **vertex** is a point in the mesh. A skin weight describes how strongly a bone influences that point. The test observed a patch of 512 weighted vertices at a fixed evaluation frame.

The same patch was recorded at baseline, half movement, peak movement, half movement on return, and restoration of the saved original pose channels. The peak command was 0.10 radians, about 5.73 degrees. These are five states in one controlled sweep, not five independent biological experiments.

Recorded state	Angle change	Vertices moved	Maximum displacement
Baseline	0	0	0
Outward halfway	+0.05 rad	512	0.0083072
Peak	+0.10 rad	512	0.0166092
Return halfway	+0.05 rad	512	0.0083072
Saved-state restore	Original channels	0	0

Displacements are in **uncalibrated asset-world units**, not millimeters. For each point, displacement is the distance between its current and baseline coordinates. The largest distance is the maximum displacement. RMS displacement combines the squared distances into one overall movement measure.

Why this was a scoped pass

All 512 sampled vertices moved at each nonzero command. The outward and return halfway coordinates matched. The final recorded coordinate differences had maximum and RMS values of exactly zero, and the original raw pose channels were restored exactly. Saved geometry also included 1,429 edges and 918 triangles.

What the visuals and this result mean

Handout 19 contains the measured geometry. Its 25-second animation displays the five saved coordinate sets; the frame durations are presentation timing, not measured breathing or response timing. The three views are point-cloud reconstructions of the same saved patch.

This test does not establish tissue forces, passive recoil, calibrated touch or nerves reaching NewBrain. Those require separate models and tests. The required future path includes skin sensors, anatomical nerve mapping, spinal/central processing, the correct avatar brain recipient and a separately verified motor return. Lung and body interactions remain future measured work.

05 / Answers you can use in the meeting

Short explanations, with enough detail to answer a follow-up.

"Did NewBrain learn something new?"

"We have not yet demonstrated a new acquisition gain. One diagnostic saved changed parameters, but its answer was already correct and the full run failed its runtime checks. The benchmark is still 2/12."

"Does the NumPy result prove learning?"

"No. The check summed the fixed integers 0 through 63. Their expected sum is $63 \times 64 / 2 = 2016$. Getting that value shows that the numerical prerequisite ran and its result was captured; it is not a learned association."

"How do you know the skin moved?"

"We saved the evaluated coordinates of the same 512 weighted points before, during and after the commanded bone movement. The three moving states changed all 512 points, and the restored coordinates and original pose channels matched exactly."

"Are these independently validated results?"

"The saved records were checked by separate project roles. We are seeking university replication and expert review; we are not presenting this as external validation or a functioning physiological patient model."

"Why has progress stopped short of a better score?"

"Several runtime, process-custody, reporting and source-integration defects have interrupted the test sequence. Some earlier attempts stopped for low available RAM, but later failures had other causes. We have not proved that RAM is the cause of the 2/12 score."

"Will more RAM or a larger GPU fix it?"

"It would give us room to evaluate larger models and future simulations. It does not guarantee learning. I am asking for research-computing access, profiling help and collaborators so we can separate software problems from measured hardware limits."

"What could the university reproduce?"

"A versioned test recipe, declared controls and tolerances, environment details, aggregate results and appropriate raw evidence can support replication. Private data and protected answer keys need an approved collaboration arrangement. Repeating the study would use new run records."

Evidence to keep with this guide

Skin: BODY-SKIN-590-CONTROL-001; qualified evidence review 630. **Teaching:** M1-002 correct-label diagnostic; failed-run evidence review 665. **Companion handouts:** 18 covers the teaching observations and resource priorities; 19 contains the measured skin views. The original evidence remains preserved privately. The current twelve-check score and Qwen status must be rechecked before each meeting.