

01 / What we can demonstrate

Kira Labs | Plain-language test methods and meeting answers | Evidence as of October 16, 2026

"We have measured controlled skin movement and exact restoration in a tested forearm patch. NewBrain has shown a recorded parameter update, but its last measured benchmark remains 2 out of 12. I am seeking university collaborators and computing access to investigate the learning gap and validate the next body-system experiments."

The current evidence

Area	Recorded result	Meaning
NewBrain benchmark	2/12: absence 2/2; facts 0/4; numerical 0/6.	No measured improvement beyond 2/12.
Numerical prerequisite	NumPy 2.4.6 ran a fixed calculation returning 2016.	A software calculation ran; this does not prove learning.
Teaching diagnostic	8,837 of 11,914 finite parameters changed; the full run failed runtime checks.	A local update was observed; new acquisition was not qualified.
Weighted skin	512 sampled vertices moved; the final recorded coordinates and raw pose restored exactly.	A scoped geometry control passed.
Lungs, nerves and Qwen	No qualified coupled lung/brain or skin/brain response; Qwen untested.	These are development and test objectives.

Three words with different meanings

Prepared: the recipe or code exists. **Executed:** an attempt ran, which may fail. **Passed:** the declared checks were met in a stated scope and the saved evidence and cleanup were reviewed.

This guide explains existing evidence and an unrun study plan. Producing this document did not run another scientific experiment.

02 / Exactly how 2/12 was scored

A strict twelve-check synthetic benchmark, with the original questions and answer key preserved.

What counts as one successful check?

The complete generated token sequence must equal the private expected sequence, and the input must contain zero unknown tokens. A **token** is an item in the model's vocabulary. The score is the number of successful checks divided by all 12 checks.

Check type	Required response	Recorded score
Four fact checks	The answer label, a synthetic source identifier and a version identifier: a complete three-token sequence.	0/4
Six numerical checks	The expected content label plus its required source and version: a complete three-token sequence.	0/6
Two absence checks	The expected single abstention token: withhold an unsupported answer.	2/2

2/12 is about 16.7% on this specific strict benchmark. It is not a general intelligence percentage, a medical accuracy estimate or an answer-label-only score.

A useful structural finding - not a new score

A case-free review found that all 12 saved baseline outputs contain one token, while 10 expected responses require three. This is a reason to investigate response length and end-of-response behavior. It does not establish the cause, show that the first answer label was correct, or justify changing the scoring rule.

How the cases were protected

The questions, expected answers, starting model and original teaching recipe were frozen and identified by file fingerprints. A fingerprint detects a changed file. Protected questions and answers stay with a separate scoring role; source reviewers use case-free descriptions. There is no answer-label regrade, dropped failure or reduced denominator.

These are known-vocabulary synthetic fact and numerical-association checks with reworded prompts and absence cases. They do not establish general arithmetic, unfamiliar-language ability, broad personal-memory recall or clinical competence.

03 / How an automatic test runs

A fixed recipe, controlled execution, preserved observations and a reviewed conclusion.

1. Define the question and rule

State what is being tested and the result that would pass. Record source versions, inputs, starting state, expected outputs, tolerances and resource limits before execution.

2. Check current readiness

Pause parallel experimental work and scheduled background work. Verify fresh run permissions, an unused attempt identifier, output locations and the cooperative stop condition.

3. Check available memory

The current NewBrain runtime requires at least 8 GiB available RAM at startup and preserves a 3 GiB reserve. Available memory is free headroom, not the amount installed in the computer.

4. Execute the declared operation

The scripts start the specified processes and record available raw returns, errors, timestamps, identities, outputs and resource observations. Each kind of test has its own checks.

5. Finish cleanup before resuming work

Keep the parallel tracks paused through shutdown, owned-process checks and evidence closure. Preserve uncertain or missing evidence as a failure or limitation; do not infer a successful exit.

6. Review the saved evidence

A separate project role checks the records against the recipe, followed by final review. A failed attempt remains saved and consumed. A later attempt requires a new identifier and fresh gates.

Other current bounds: a 40-second child / 45-second whole-run deadline, a bounded process family and roughly 819 MiB family memory ceiling. The selected short private/outside phases are 20/22 seconds. Configured numerical-provider capacity is two; exact active numerical-worker count is unmeasured. Saved-report limits concern evidence bytes, not RAM. Sampled resource observations are not continuous measurements.

A syntax check can show that source text parses. A startup, cleanup, numerical-library or report-transport check can qualify that engineering operation. None of these automatically establishes a learning gain. New source versions need their own applicable controls.

04 / What the learning diagnostic showed

Training behavior, answer accuracy and whole-run qualification are separate measurements.

One actual update, on an already-correct answer

The diagnostic saved the model before and after one training update. Review compared its adjustable numbers, generated answer, probability and training objective. The requested answer was already correct before teaching.

Measurement	Before	After
Generated answer	Already correct	Same answer
Correct-answer probability	0.9991045212	0.9990638723
Total training objective	7.8074396472	7.5016123944
Finite parameter values changed	Compared with starting state	8,837 of 11,914

A **parameter** is an adjustable model number. The **objective** is the training formula being minimized. Here the objective decreased while correct-answer probability decreased slightly. Changed parameters show that an update occurred; they do not show a newly learned fact or a better benchmark score.

Why the complete run was rejected

The process review lacked retained evidence proving the completion of one original controller generation. Some final worker exports were missing. The attempt failed its runtime checks. Its local observations remain preserved, but it has no qualified acquisition, retention, generalization or new twelve-check score. This was not another low-RAM startup stop.

The current faithful twelve-check study - unrun

The plan returns to the original decoder component, twelve questions, strict answer key and original complete teaching targets. It compares baseline with post-teaching responses. The fixed teaching schedule is four bursts of 16 updates: **64 updates and 384 example exposures**, using repeated original teaching objects. An exposure is not a new unique fact.

The source recipe is prepared; matching process controls and the actual study are still pending. There is no intermediate protected scoring or tuning. The original score remains 2/12. The older proposed six-case study does not replace this twelve-check comparison.

New acquisition requires a target that was incorrect before teaching to become correct afterward. Fresh-process persistence, retention of older learning and generalization need separately declared evidence. Qwen needs its own qualified comparison.

05 / How the skin result was measured

One controlled forearm sweep, with the same 512 weighted mesh points recorded at five states.

The question and the setup

The test asked whether moving a forearm bone moves its weighted skin and whether the original state restores. A **vertex** is a point in the mesh; a **skin weight** describes a bone's influence on that point. The same patch was evaluated at a fixed frame. Four bone matrices were also recorded.

Recorded state	Angle change	Points moved	Maximum distance
Baseline	0	0	0
Outward halfway	+0.05 radians	512	0.0083072
Peak	+0.10 radians	512	0.0166092
Return halfway	+0.05 radians	512	0.0083072
Restore saved pose	Original channels	0	0

The peak command was about 5.73 degrees. Distances are **uncalibrated asset-world units**, not millimeters. These are five recorded states of one sweep, not five independent biological experiments.

How the numbers were calculated

For each vertex, subtract its baseline position from its current position and calculate the straight-line distance. The largest distance is the maximum displacement. **RMS** is the square root of the mean of all squared distances: an overall movement measure. At peak, RMS displacement was 0.0109012 asset-world units.

Why this was a scoped pass

All 512 sampled vertices moved at every nonzero command. Outward and return halfway coordinates matched. Restored coordinates had maximum and RMS displacement exactly zero in the saved data; the original raw pose channels restored exactly. The saved patch included 1,429 edges and 918 triangles.

Limits: this does not prove tissue forces, passive recoil, collision/contact, appearance realism or touch arriving at NewBrain. Skin-to-brain nerves and the correct per-avatar recipient need separate tests. Coupled lung, blood, brain and other organ behavior remains unvalidated.

Handout 19 includes three views and a 25-second display of the five recorded coordinate sets. Its frame durations are presentation timing; it is not a continuous physics or lung-engine recording.

06 / Answers for the meeting

Short answers you can use, with a clear next research step.

Did NewBrain learn something new?

"We have recorded a parameter update, but not a qualified new-acquisition gain. The last benchmark is still 2/12 under its original strict full-response rule."

Does NumPy 2.4.6 prove learning?

"No. The fixed integers 0 through 63 sum to $63 \times 64 / 2 = 2016$. That result shows the numerical prerequisite ran; it is not a learned association."

Why is there no better score yet?

"Earlier attempts included low available RAM. Later attempts encountered software integration, process-completion and reporting defects. We have not established RAM as the cause of the 2/12 learning score."

Will more RAM or a larger GPU solve it?

"They provide headroom for larger comparisons and future simulations, but do not guarantee learning. We need profiling, controlled experiments and collaborators to separate hardware limits from software and model limitations."

Are the results independently validated?

"Separate roles within the project reviewed saved evidence. That is not external university replication, peer review or clinical validation. I would welcome a university reproducing the controlled tests."

What would the university receive?

"A versioned recipe, environment details, fixed scoring and tolerances, aggregate results, and appropriate raw evidence. Private records and protected keys require an agreed access arrangement. A replication uses fresh attempt records."

Evidence references: original twelve-check case-free review 676 and recipe descriptor 677; teaching diagnostic M1-002 / failed-run review 665; weighted-skin control BODY-SKIN-590-CONTROL-001 / qualified review 630. Handout 19 contains the skin views. Original evidence and failed attempts remain preserved privately. Recheck the score and Qwen status before each meeting.