NEO AI MODEL LAB - OFFICIAL TEST NO. 4 - REPLICATION

Second run, same verdict: top-model accuracy at a lower cost, announced in advance and met twice

The same 99 tasks, the same two targets, a new day. A result that holds when it is run again is worth more than a result seen once.

198 of 198

tasks solved exactly by NEO AUTO over the two runs, level with Claude Opus 5.5 and GPT-5.5

38 to 52%

lower cost per successful task than the cheaper reference, over the two runs; published target: 30%

2 of 2, twice

targets published and time-stamped before each run, met on both days

Official test report
NEO AI - a My NEO Group company
Technology developed by NEO TECH LABS TECHNOLOGY LLC

23 November 2026
Public edition

THE SECOND RUN

The same test, run again: both targets met a second time

Official Test No. 3 announced two targets in advance and met them. One day later, Official Test No. 4 asks the only question that matters after a good result: does it hold when the test is run again?

The protocol was published and time-stamped by two outside authorities at 12:24 UTC on 23 November 2026; the first answer was requested at 12:26:11 UTC. Same five systems, same 99 tasks with one known answer, same two targets, live calls at real prices, no interruption.

THE KEY RESULT

NEO AUTO again solved every one of the 99 tasks, exactly as Claude Opus 5.5 and GPT-5.5 did, at 0.0263 USD per successful task against 0.0426 USD and 0.0631 USD. Both published targets are met a second time.

The second run at a glance

FACT - MEASURED

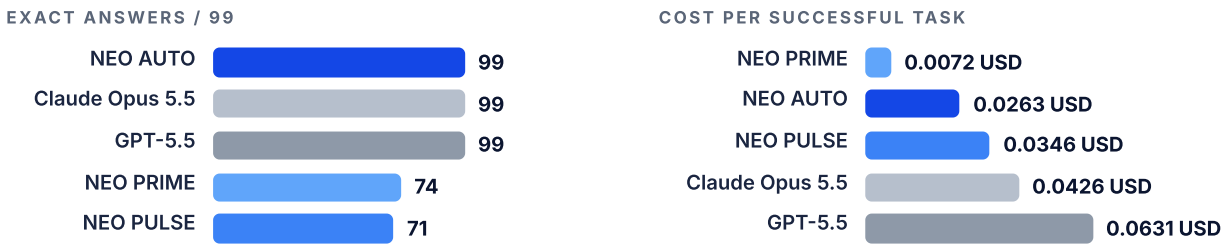
99 / 99
NEO AUTO exact answers: every task solved, again

38%
lower cost per successful task than the cheaper reference; target 30%

0 point
of accuracy below the best reference; target at most 7

495
live answers across five systems, sealed the second the run ended

The five systems



The two targets, published before the run

FACT - TIME-STAMPED

TARGET	PUBLISHED	MEASURED	OUTCOME
Cost per successful task of NEO AUTO, against the cheaper of the two references	at least 30% lower	**38.3% lower**	**Met**
Accuracy of NEO AUTO, against the better of the two references	at most 7 points below	**0 point below**	**Met**

An answer counts only if its final answer equals the expected one: no partial credit, no reviewer. The cost per successful task is the cost of all 99 answers of a system divided by the number of tasks it solved.

TEST 4 AGAINST TEST 3

What held on the second day, and what moved

Two runs of the same protocol, one day apart. Set side by side, they separate what is a property of LILI from what belongs to a single day.

System by system

FACT - MEASURED

SYSTEM	EXACT, TEST 3	EXACT, TEST 4	COST PER SUCCESS, TEST 3	COST PER SUCCESS, TEST 4	MEDIAN TIME, TEST 3	MEDIAN TIME, TEST 4
NEO AUTO	99 / 99	**99 / 99**	0.0213 USD	**0.0263 USD**	8.5 s	**10.9 s**
Claude Opus 5.5	99 / 99	**99 / 99**	0.0448 USD	**0.0426 USD**	9.4 s	**8.9 s**
GPT-5.5	99 / 99	**99 / 99**	0.0611 USD	**0.0631 USD**	12.9 s	**12.1 s**
NEO PRIME	66 / 99	**74 / 99**	0.0081 USD	**0.0072 USD**	8.7 s	**9.3 s**
NEO PULSE	69 / 99	**71 / 99**	0.0402 USD	**0.0346 USD**	8.1 s	**7.6 s**

198 / 198
NEO AUTO exact answers over the two runs: not one task missed

52% and 38%
cost advantage on each day, against a target of 30% announced in advance

99 of 99
NEO AUTO answers checked by an independent model before they were given, as in Test 3

What held

INTERPRETATION

- **Accuracy held completely.** NEO AUTO solved every task on both days, level with the two leading models. The verification layer did its work each time: no wrong answer was returned.
- **The cost target held with a margin.** The advantage over the cheaper reference was 52% on the first day and 38% on the second. The honest statement of LILI's advantage on these tasks is therefore a range, 38 to 52%, and its floor is above the target.
- **The lighter profiles improved.** NEO PRIME and NEO PULSE each solved more tasks than on the first day, and NEO PRIME remains the lowest-cost system of the five.

What moved

FACT, THEN INTERPRETATION

- **The cost advantage was smaller on the second day.** LILI's verifier asked for a stronger answer on 24 tasks, more often than on the first day. Each of those tasks ended exact: LILI spent more where its verifier was not satisfied, and accuracy was never traded for cost.
- **Time did not repeat.** NEO AUTO's median answer took 10.9 s against 8.9 s for Claude Opus 5.5; on the first day NEO AUTO had been the quicker of the two. The two targets concern cost and accuracy; this note makes no claim on time.
- **This is what a repetition is for.** A single run gives a figure; two runs give a range and show which results can be relied upon. Accuracy and the cost target can. The engineering work that follows is on the stability of routing, so that the advantage of the best day becomes the advantage of every day.

PROOF

The same chain of proof, a second time

The second run carries the same chain of proof as the first: rules time-stamped before the run, results sealed when it ended, an evidence pack that anyone can recompute without NEO AI.

2 min
between the third-party time-stamp of the rules and the first answer

13:29:26
UTC: results sealed and time-stamped by DigiCert, the second the run ended

495
per-task results published: system, task, status, time, cost, exact or not

0
discrepancy when the public verifier and a second, independent program recompute the figures

The register of proofs of Official Test No. 4

FACT - RECORDED

PROOF	WHAT IT ATTESTS	STATE
DigiCert - RFC 3161 time-stamp of the protocol	The rules and the targets existed at 12:24:15 UTC, before the run	Verified
FreeTSA - RFC 3161 time-stamp of the protocol	The same date, attested by a second, independent authority	Verified
Sigstore - signature of the protocol	Who published the rules, in a public log that cannot be edited	Verified
Proof Ledger - entry no. 3,669	The campaign is sealed in the hash chain; the entry recomputes to its hash	Verified
DigiCert - RFC 3161 time-stamp of the seal	The results existed at 13:29:26 UTC, the second the run ended	Verified
Merkle root of the results - RFC 6962	495 public rows bound by one root: none can be changed, added or removed	Verified
Manifest fingerprint - RFC 8785	Files, figures and limits of the evidence pack stand or fall together	Verified
Sigstore - signature of the evidence pack	Who sealed the pack, in the public Rekor log	Verified
FreeTSA - time-stamp of the evidence pack	The pack existed before the round of the public draw	Verified
OpenTimestamps - Bitcoin	A date committed to the Bitcoin block chain, which no authority can withdraw	In progress
drand - League of Entropy	Round 34,217,812 designates the ten tasks published in full	In progress
Hugging Face - open dataset	The per-task results, read back file by file against the pack	Verified
Public repository - full history	Protocol, pack, verifier and workflow are public: github.com/myneogroup-ai/neolabs-proofs	Verified

Permanent archives. The per-task results are an open dataset: huggingface.co/datasets/NEOTECHLABS/neo-ai-evaluations-test-4. The edition is also deposited with the Internet Archive, Kaggle and Zenodo at CERN; each deposit is recorded in the public register as it is confirmed.

Recompute it yourself

```
git clone https://github.com/myneogroup-ai/neolabs-proofs
cd neolabs-proofs
node verify/cli.mjs verify packs/test-4
python3 recalc/recalculate.py packs/test-4
```

Free software, no account: two programs that share no code recompute every figure.

Verify the seal

```
openssl ts -verify -in test-4-ledger-3669.tsr -CAfile
digicert-trusted-root-g4.pem
  -digest cb87417941d05e24f97749f9cdf9626d2ad6980ad9ae
358639c655409be5439b
```

One standard command confirms the DigiCert token of the seal.

SUMMARY

What the repetition shows

Two official tests in two days on the same ground, both announced in advance, both sealed, both public. LILI's central promise now rests on a result that has been repeated.

198 / 198
every task solved on both days, level with Claude Opus 5.5 and GPT-5.5 (NEO AUTO)

38 to 52%
lower cost per successful task than the cheaper reference, over the two runs

4 / 4
targets published before the runs and met: two on each day

990
per-task results published over the two runs, each one recomputable by anyone

Accuracy is repeatable

NEO AUTO answers 99 of 99 tasks exactly on both days, on reasoning, data and code. Every answer was checked by an independent model before it was given.

The cost advantage has a floor

52% on the first day, 38% on the second, against a published target of 30%. The floor of the range, not its best day, is what LILI commits to.

Three profiles, one brain

NEO PRIME remains the lowest-cost system of the five; NEO PRIME and NEO PULSE both solved more tasks than on the first day.

Published whatever the figure

The protocol was time-stamped before the run, so the result could not be chosen afterwards. This note states what moved as plainly as what held.



SEEN, NOT ONLY MEASURED

LILI's brain, recorded live during the test

While LILI worked, the platform recorded its brain in full: every layer, every decision, every call. Two of those recordings are replayed at real speed: a NEO AUTO task and a NEO PULSE task of this test. Nothing is staged: the films show what LILI actually did. The AI services appear under neutral names. Watch them on lilineo.com/evaluations/official-tests/test-4/.

This was LILI's fourth official test, and its first repetition. It confirms what the third test announced: the accuracy of the leading models, at a cost lower than theirs by more than the target published in advance. Each official test, favourable in every figure or not, is published with the same rigour and the same proof.

ANNOUNCED
REPEATED
PROVEN

Important notice. This note presents the results of an internal NEO AI test run on 23 November 2026 on 99 tasks with one known answer and five systems, with each LILI profile as configured on that date. It repeats Official Test No. 3 of 22 November 2026 on the same tasks with the same targets: the tasks and the behaviour of the systems on them were known before the run, and ten of the tasks had been published in full. Costs are computed by the platform from the token counts reported by the providers and the price tables in force; they are not invoices. Each task was answered once by each system on each day; no confidence interval is claimed. The three leading systems solved every task: the test measures a difference of cost at equal accuracy on this set, not a difference of accuracy. NEO AUTO answered more slowly than one reference in this run. Nine answers of the two lighter profiles ended in an error and are counted as not exact; 55 answers of NEO PULSE were priced at the highest known price of their provider. The third-party time-stamps are standard RFC 3161 time-stamps: they certify the existence and integrity of the recorded files at the stated times. The authorities and services named are independent providers of public services and have not reviewed this note. The test has not been audited or reproduced by an external party, and results may differ on other tasks, dates or configurations. This document is provided for information only. Claude Opus 5.5 and GPT-5.5 are products of their respective owners, used as references through their public interfaces.