NEO AI MODEL LAB - OFFICIAL TEST NO. 3

Top-model accuracy at less than half the cost: announced in advance, sealed and proven

Same tasks, real calls, real prices. This time the rules and the targets were published and time-stamped before the first answer, and anyone can recompute the result.

99 of 99

tasks solved exactly by NEO AUTO, level with Claude Opus 5.5 and GPT-5.5

2.1x

lower cost per successful task than Claude Opus 5.5; 2.9x lower than GPT-5.5

2 of 2

targets published before the run, both met; twelve proofs anyone can check

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

22 November 2026
Public edition

INTRODUCTION

From a certified measurement to a result announced in advance

On 18 November 2026 the Model Lab measured quality, speed and cost, and an outside authority certified when the results existed. Four days later, this third test adds the two things that turn a certified measurement into a commitment kept.

First, **the result was announced in advance**: the rules of the test and two numerical targets were published and time-stamped by outside authorities two minutes before the first answer was requested. Second, **accuracy is no longer a matter of opinion**: each of the 99 tasks has one known answer, so an answer is exact or it is not, and anyone can recompute every figure of this report from the public evidence.

LILI met both targets with room to spare: it now delivers the accuracy of the leading models at less than half their cost, and proves it in public.

THE KEY RESULT

On 99 reasoning, data and code tasks, NEO AUTO solved every task, exactly as Claude Opus 5.5 and GPT-5.5 did, at 0.021 USD per successful task against 0.045 USD and 0.061 USD, and in a median 8.5 s against 9.4 s and 12.9 s.

The third test at a glance

FACT - MEASURED

99 / 99
NEO AUTO exact answers: every task solved

2.1x
lower cost per successful task than Claude Opus 5.5

52%
cost reduction measured, against a published target of 30%

495
live answers across five systems, none interrupted

NEO AUTO against the two references



16x
lower cost on the typical task: NEO AUTO's median answer costs 0.0023 USD, against 0.0370 USD for Claude Opus 5.5

8.5 s
median time to a complete answer: the fastest of the three leading systems

2 min
between the third-party time-stamp of the rules and the first answer: the targets could not be moved

WHY LILI WINS

An intelligence architecture, not another model

The leading AI labs spend billions of dollars to train ever larger models. NEO AI does not enter that race, and does not need to. LILI is a layer of intelligence built **above** the best models of every provider: it decides how much compute each request deserves, verifies the result, remembers, and proves what it did.

A FRONTIER MODEL LAB

Builds one model

Billions of dollars of compute and years of training. Every request, simple or hard, is answered by the same large model at the same price.

A SIMPLE ORCHESTRATOR

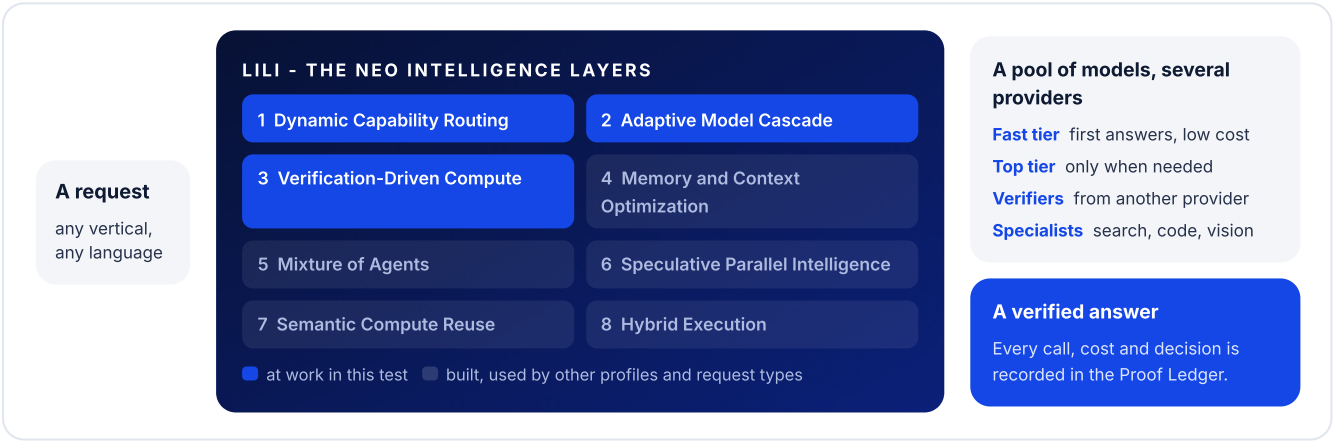
Connects APIs

Forwards each request to a model by fixed rules. Easy to build and easy to copy: it adds no judgement, no verification, no memory and no proof.

LILI BY NEO AI

Decides, verifies, remembers, proves

Eight coordinated layers choose the right compute for each request, check every answer with an independent model, learn through a memory graph and record each call. Every new model on the market makes LILI stronger.



The eight NEO layers

3 OF 8 AT WORK IN THIS TEST

1. **Dynamic Capability Routing**
   Reads each request and selects the tier, the tools and the data it needs, before any model is paid.
2. **Adaptive Model Cascade**
   A fast, low-cost model answers first. A top-tier model is called only when the answer needs it.
3. **Verification-Driven Compute**
   An independent model from another provider checks the answer. Evidence decides whether more compute is spent.
4. **Memory and Context Optimization**
   LILI sends the facts that matter from its memory graph and a running summary, instead of the whole history.
5. **Mixture of Agents**
   For expert work: up to three models from different providers answer in parallel, then one synthesis and one critic.
6. **Speculative Parallel Intelligence**
   Two approaches are sketched in parallel and scored blind; the stronger one is developed.
7. **Semantic Compute Reuse**
   A verified answer to an equivalent question is reused for the same person instead of being computed again.
8. **Hybrid Execution**
   Designed to place models on API, private or self-hosted capacity. Today every model runs through APIs.

The goal is intelligence per compute: the accuracy of the best model for a fraction of its cost. One LILI serves every NEO AI vertical through one gateway, so each improvement benefits all of them at once. This test measures the result with three layers only: every task solved, for less than half the cost. Memory was deliberately left out, so the comparison is on strictly equal terms.

THE INSTRUMENT

How the architecture is put to the test

The Model Lab runs LILI and reference models on identical tasks, through live calls, and records everything. For this third test the bar was raised twice: the answer is known in advance, and so was the target.

1 **Same tasks**
99 original tasks with one known answer, identical for every system

2 **Targets first**
Rules and two targets published and time-stamped before the run

3 **Real calls**
495 live answers, real providers, real prices

4 **No opinion**
An answer is exact or it is not: an automatic comparison, no reviewer

5 **Proof Ledger**
Every run sealed in an append-only, hash-chained register

What was tested

SYSTEM	WHAT IT IS	HOW IT WAS CALLED
NEO AUTO	LILI, balanced profile: a first answer, an independent verification, and a stronger model only when needed	Through LILI
NEO PULSE	LILI, speed profile: the fastest route to an answer	Through LILI
NEO PRIME	LILI, economy profile: the lowest cost per answer	Through LILI
Claude Opus 5.5	Reference: a leading single model	One call, no orchestration
GPT-5.5	Reference: a leading single model	One call, no orchestration

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	**52.4% lower**	**Met**
Accuracy of NEO AUTO, against the better of the two references	at most 7 points below	**0 points below**	**Met**

What changed since the second test

THREE IMPROVEMENTS

The result is announced first

The protocol, the 99 tasks by their fingerprint and the two targets were time-stamped by two outside authorities at 08:13:37 UTC. The first answer was requested at 08:15:25 UTC.

Accuracy without opinion

Each task has one expected answer, computed by a program and checked by a second one. No reviewer, human or AI, has anything to judge.

Anyone can recompute it

The per-task results are public. A free verifier recomputes every figure and the outcome of the two targets, and fails on any difference.

Scope: 99 original tasks in English and French - 36 of reasoning, 36 of data, 27 of code - each answered once by each of the five systems: 495 answers, in new conversations with no history and no memory.

RESULTS - ACCURACY

Every task solved: level with the two leading models

Ninety-nine tasks with one right answer each: multi-step reasoning, data computations, code to read and trace. NEO AUTO answered every one of them exactly.

Exact answers out of 99 FACT - MEASURED



An answer counts only if its final answer equals the expected one. No partial credit, no reviewer.

SYSTEM	EXACT ANSWERS	REASONING	DATA	CODE	BEST SUITED TO
NEO AUTO	**99 / 99**	36 / 36	36 / 36	27 / 27	Demanding work where the answer must be right
Claude Opus 5.5	**99 / 99**	36 / 36	36 / 36	27 / 27	Reference
GPT-5.5	**99 / 99**	36 / 36	36 / 36	27 / 27	Reference
NEO PULSE	**69 / 99**	13 / 36	31 / 36	25 / 27	Instant replies, data and code at speed
NEO PRIME	**66 / 99**	13 / 36	29 / 36	24 / 27	Very high volume at minimal cost

99 = 99 = 99
NEO AUTO level with the two leading models: the ceiling of the test, reached

99 of 99
NEO AUTO answers checked by an independent model before they were given

56 and 53 / 63
NEO PULSE and NEO PRIME on data and code, in seconds and for a fraction of a cent

How to read this INTERPRETATION

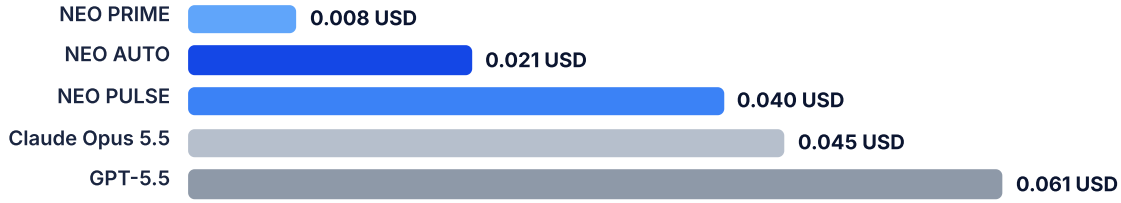
- **NEO AUTO delivers the headline.** It reaches the accuracy of the two leading models on every class of task. On this ground nothing is left to gain on accuracy: the difference is made on cost and time, the two figures a customer pays for.
- **Verification is what makes it safe.** On 16 tasks LILI's own verifier asked for a stronger answer before anything was returned, and every one of those tasks ended exact. LILI does not hope the cheap answer is right: it checks.
- **NEO PULSE and NEO PRIME are built for another job.** They handle data and code at speed and at minimal cost. Multi-step reasoning is exactly where LILI routes work to NEO AUTO, which is why one brain carries three profiles.

RESULTS - COST AND SPEED

Less than half the cost, and faster, at equal accuracy

Cost per successful task

FACT - MEASURED



The cost of all 99 answers of a system, divided by the number of tasks it solved. A failed answer still costs: it is counted.

Median time to a complete answer

FACT - MEASURED



LILI PROFILE	COST VS CLAUDE OPUS 5.5	COST VS GPT-5.5	TIME VS CLAUDE OPUS 5.5	TIME VS GPT-5.5
NEO AUTO	2.1x lower cost per successful task	2.9x lower cost per successful task	9% shorter time	34% shorter time
NEO PRIME	5.5x lower cost per successful task	7.5x lower cost per successful task	7% shorter time	33% shorter time
NEO PULSE	1.1x lower cost per successful task	1.5x lower cost per successful task	13% shorter time	37% shorter time

16x / 26x
lower cost on the typical task than Claude Opus 5.5 and GPT-5.5 (median cost of an answer)

5.5x
NEO PRIME against the cheaper reference: the lowest cost per successful task of the five systems

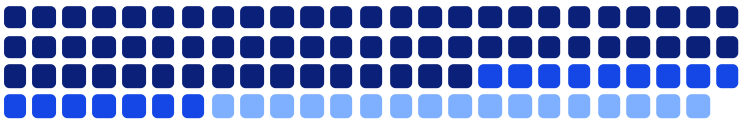
8.1 s
NEO PULSE: the shortest time to a complete answer of the five systems

Ratios of the costs per successful task and of the median times above. The advantage of NEO AUTO is stated on the average, the more demanding figure: it includes every task that LILI chose to send to its top tier. Each profile is built for a different job; page 9 sets them side by side.

ANATOMY OF THE WHOLE TEST

The same mechanism, across all 99 tasks

The register holds the path of every answer, so the mechanism can be counted. Each square is one of the 99 tasks answered by NEO AUTO.



- 66 tasks: fast-tier answer, verified and accepted
- 16 tasks: escalated to the top tier by LILI's verifier
- 17 tasks: sent directly to the top tier

What LILI decided, task by task

FACT - RECORDED

WHAT LILI DID	TASKS	TYPICAL COST PER ANSWER	TYPICAL TIME	EXACT
First answer by a fast-tier model, verified by an independent model and accepted; top-tier call avoided	66	0.0022 USD	9.0 s	66 / 66
Routed directly to a top-tier model, then verified and accepted	17	0.0490 USD	11.2 s	17 / 17
First answer by a fast-tier model, not accepted by the verifier; a top-tier model answered	16	0.0708 USD	25.1 s	16 / 16
NEO AUTO, 99 answers	**99**	**0.0213 USD average**	**8.5 s median**	**99 / 99**
Claude Opus 5.5, 99 answers	99	0.0448 USD average	9.4 s median	99 / 99

21x
lower cost than a reference answer on the 66 tasks LILI settled at the fast tier, each one verified and exact

66 of 99
top-tier calls that LILI established it did not need to make

16 of 16
escalations decided by the verifier that ended with the exact answer

What this shows

INTERPRETATION

- **The saving comes from calls that were not made.** On 66 of 99 tasks LILI established, by verification, that a fast-tier answer was right, and did not buy a top-tier one. Those answers cost about 0.002 USD each.
- **The top tier is used when it earns its cost.** LILI called it on 33 of 99 tasks: 17 recognised as hard at first reading, 16 after its verifier declined the first answer. Every one of them is counted in the averages of this report: the 2.1x advantage is net of all escalations.
- **Accuracy is bought only where it is needed.** A single model pays its full price on every task. LILI pays it on one task in three, and reaches the same 99 out of 99.
- **The next gain is already identified.** Recognising a hard task at first reading, instead of after a first attempt, removes the cost and the time of that attempt. It is the next engineering step.

Paths, typical costs and typical times are read from the platform's own record of each NEO AUTO answer. Models are named by tier and role; the exact model and provider of every call are recorded in the Proof Ledger and available on request under confidentiality.

THREE PROFILES, ONE BRAIN

The right compute for each request

LILI does not sell one model. It sells the choice of the right amount of compute for each request. This test shows the three profiles side by side, on the same demanding tasks.

NEO AUTO - BALANCED

99 / 99

Every task solved at 0.021 USD per successful task. The profile for work where the answer must be right: verified every time, top tier only on evidence.

NEO PRIME - ECONOMY

0.008 USD

The lowest cost per successful task of the five systems, 5.5x below the cheaper reference. 53 of 63 data and code tasks solved. Built for very high volume.

NEO PULSE - SPEED

8.1 s

The shortest time to a complete answer of the five systems. 56 of 63 data and code tasks solved. Built for immediate replies.

The five systems, one table

FACT - MEASURED

SYSTEM	EXACT	DATA AND CODE	REASONING	COST PER SUCCESSFUL TASK	MEDIAN TIME	BEST SUITED TO
NEO AUTO	99 / 99	63 / 63	36 / 36	0.021 USD	8.5 s	Demanding work, right every time
NEO PRIME	66 / 99	53 / 63	13 / 36	0.008 USD	8.7 s	High volume at minimal cost
NEO PULSE	69 / 99	56 / 63	13 / 36	0.040 USD	8.1 s	Instant replies
Claude Opus 5.5	99 / 99	63 / 63	36 / 36	0.045 USD	9.4 s	Reference
GPT-5.5	99 / 99	63 / 63	36 / 36	0.061 USD	12.9 s	Reference

How to read this

INTERPRETATION

- **Each profile wins on its own ground.** NEO AUTO on accuracy at lower cost, NEO PRIME on price, NEO PULSE on time. No single model in the test leads on more than one of the three.
- **The routing rule writes itself.** Multi-step reasoning goes to NEO AUTO; data lookups, code reading and high-volume requests go to NEO PRIME and NEO PULSE. That rule is what LILI's first layer applies on every request, in every vertical.
- **The next step is precise.** The test shows exactly where the two lighter profiles gain most: multi-step reasoning. Sending those requests upward automatically is a routing change, not a new model.

WHAT THIS MEANS FOR A CUSTOMER

A customer who needs the right answer gets the accuracy of the best models for less than half their price. A customer who needs volume or speed gets it for a fraction of a cent. Each request receives the compute it needs, and no more: one brain, three profiles, the right compute every time.

EVIDENCE

How every figure was calculated, proven and recorded

No figure in this report was typed by a person. Each one is computed from records written by the platform while the test ran, and can be recomputed by anyone from the public evidence.

1 — Frozen task set
99 tasks with their expected answers, version-locked and fingerprinted in the protocol before the run.

2 — Targets published
The protocol and its two targets are time-stamped by outside authorities before the first call.

3 — Live production call
Each task is sent as a real request: through LILI for the three profiles, directly to the model for the two references.

4 — Cost ledger line
Each call is priced at once: tokens reported by the provider x the price table in force. Verification calls included.

5 — Timing events
The server timestamps the start, the first text and the final answer of every run.

6 — Automatic comparison
The final answer is compared with the expected one. Exact, or not: nobody scores anything.

7 — Proof Ledger seal
The completed campaign is appended to the hash-chained register and time-stamped in the same second.

8 — Public recomputation
A free verifier recomputes every figure and the outcome of the targets from the published rows.

Accuracy

An answer is a success when its final answer equals the expected one. Reasoning, data and code tasks alike; no partial credit.

Cost

Tokens reported by the provider for each call, multiplied by the price table in force on the day, including every internal call of LILI. Divided by the tasks solved.

Time

Server timestamps at the start of each request and at the final answer. The reported time is the median over the 99 tasks.

The Proof Ledger: why the record cannot be rewritten FACT

Entry 3133
`hash fa229c4d80...`
`prev linked`

Entry 3134 - Test no. 3
`495 answers sealed`
`hash ca0fe093d4...`
`prev fa229c4d80...`

Next entry
`prev ca0fe093d4...`
`link intact`

If one figure changed
its fingerprint would no longer match, every later link would break, and the verification would report the exact entry.

Each entry contains the fingerprint (SHA-256 hash) of the entry before it. Alteration is detectable by design.

- **Append-only.** The database itself refuses any modification or deletion of an entry. Entries can only be added.
- **Sealed as a whole.** When the campaign ended, the fingerprint of all 495 answers was sealed in one entry (no. 3,134), chained to the 3,133 entries before it.
- **Bound row by row.** The public results carry a Merkle root over the 495 rows: no row can be changed, added or removed without changing it.
- **Certified in time by third parties.** The rules before the run, the seal when it ended, the evidence pack afterwards. The next two pages set out who attests what.

THIRD-PARTY PROOF

Outside authorities certify the rules before the run and the results after it

The second test proved when its results existed. This one proves two dates: the rules existed before the run, and the results the second it ended. Neither rests on NEO AI's own clock.

Rules published

Protocol, tasks by fingerprint and the two targets. Time-stamped at 08:13:37 UTC by FreeTSA and DigiCert.

Campaign runs

First answer at 08:15:25 UTC, last at 09:22:15 UTC. 495 answers, no interruption.

Only the hash leaves

32 bytes are sent to the authority. No question, no answer, no figure, no name ever leaves NEO AI.

DigiCert signs

The authority signs the hash with its own clock: 22 November 2026, 09:22:16 UTC. Serial d8658f73...aa83

The seal and its time-stamp ran automatically, in the same second the campaign ended. No person intervened.

Who attests

DigiCert is one of the world's long-established public certificate authorities, trusted by default by operating systems and browsers. Its service follows RFC 3161, the international standard for trusted time-stamps. **FreeTSA** is a second, independent time-stamp authority. **Sigstore** records who signed in a public transparency log that cannot be edited. **The League of Entropy** publishes public randomness that nobody controls.

What the token says

Signed by	DigiCert SHA256 RSA4096 Timestamp Responder 2026 1
Time certified	22 November 2026, 09:22:16 UTC
Token serial	`d8658f7366fa3662fc693602107baa83`
What was signed	Ledger entry no. 3,134, hash `ca0fe093...31ec`
Signature check	**Valid**

What this guarantees

The targets could not be moved

They carry a date earlier than every result. A changed protocol would no longer match its time-stamped fingerprint.

The results cannot be changed afterwards, even by us

Any change to an answer, a status or a cost changes the fingerprint. The token would no longer match, and anyone could see it.

The whole history is covered

Each ledger entry contains the fingerprint of the one before, so one token covers all 3,134 entries up to this test.

Nothing confidential left the company

The authority received a 32-byte fingerprint. It cannot read it back into data, and it saw no question, answer or name.

Anyone can verify it, without NEO AI

```
openssl ts -verify -in test-3-ledger-3134.tsr -CAfile digicert-trusted-root-g4.pem
  -digest ca0fe093d4255d6dbcef449f54e578eceffac9f61b200907223dee60fb3f31ec
```

One standard command, with free software and DigiCert's public root certificate, confirms the token; the token file is public, in the evidence pack. The time-stamps certify that the rules and the results existed, unchanged, at those times; the soundness of the method is what the rest of this report sets out.

PUBLIC VERIFICATION

Twelve proofs, and a result anyone can recompute

A certificate says when. This page is about what: the evidence pack of this test is public, and a free verifier recomputes the result. The reader does not have to take NEO AI's word for the arithmetic.

12
proofs attached to this test, each one checkable by anyone

4
outside parties: DigiCert, FreeTSA, Sigstore, the League of Entropy; Bitcoin as a fifth anchor

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

0
discrepancy when the public verifier recomputes the published figures

The register of proofs

FACT - RECORDED

PROOF	WHAT IT ATTESTS	STATE
DigiCert - RFC 3161 time-stamp of the protocol	The rules and the targets existed at 08:13:37 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. 3,134	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 09:22:16 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	Verified
drand - League of Entropy	Round 34,185,412 drew the ten tasks now published in full	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 also published as an open dataset: huggingface.co/datasets/NEOTECHLABS/neo-ai-evaluations-test-3. The published release is archived by Zenodo at CERN under DOI 10.5281/zenodo.24452513, and the archive of the repository has been requested from Software Heritage.

The public draw

The texts of the tasks stay private so the set remains usable. To show they are real, ten of them are published in full - chosen not by NEO AI but by a public source of randomness, round 34,185,412 of the drand beacon, fixed in the protocol before the run. The evidence pack was sealed before that round existed.

Recompute it yourself

```
git clone https://github.com/myneogroup-ai/neolabs-proofs
cd neolabs-proofs
node verify/cli.mjs verify packs/test-3
```

Free software, no account, no network for the check itself. It recomputes the exact answers, the cost per successful task and the outcome of the two targets.

MOMENTUM

From a certified result to a commitment kept in public

Four days of progress

FACT - READ FROM THE LEDGER

Tasks in the test

TEST 2 **34**

TEST 3

99

How accuracy is established

TEST 2 **reviewers**

TEST 3

known answer

Rules and targets

TEST 2 **after the run**

TEST 3

before the run

Proofs anyone can check

TEST 2 **1 at the time**

TEST 3

12

The third test is also the more demanding one for LILI. The tasks have a single right answer, so there is no credit for a good attempt, and the references are at their strongest: both score 99 out of 99. On this ground LILI sends one task in three to its top tier - and still costs less than half. In the second test, on analysis work, the same architecture cost about one fifth of the reference. **Two families of tasks, two independent measures, one conclusion: the saving is structural.**

	TEST NO. 2 - 18 NOVEMBER	TEST NO. 3 - 22 NOVEMBER	WHAT IT ADDS
Ground	34 analysis tasks, judged out of 10	99 reasoning, data and code tasks, one right answer	A second family of tasks, harder to argue with
NEO AUTO	8.70 / 10 against 8.91; 4.9x lower cost per answer	99 / 99 against 99 / 99; 2.1x lower cost per successful task	Parity with the best model confirmed, this time exactly
Proof	Sealed and time-stamped by DigiCert when it ended	Targets time-stamped before the run; pack public; twelve proofs	From "certified afterwards" to "announced in advance"
Verification	Token and ledger entry on request	Public repository, free verifier, public draw	From "ask us" to "check it yourself"

The next milestones

ALREADY BUILT INTO THE MODEL LAB

HEIGHT

Tasks beyond the ceiling

A set on which the leading models themselves fail part of the time, to measure accuracy beyond the ceiling LILI reached here.

DEPTH

Long analyses

The profiles built for expert work, on long documents, where the five other NEO layers come into play.

PERMANENCE

Permanent archives

The open dataset and the DOI are live. Next: a software archive for every evidence pack, as a standard step of each test.

INDEPENDENCE

External replication

Date and integrity are attested by third parties; next, the test reproduced by an outside party.

Path of proof

Test no. 2 - 18 Nov

Quality measured. Results sealed and time-stamped by a third party.

Test no. 3 - 22 Nov

Targets announced in advance and met. Public evidence, twelve proofs.

Next tests

Harder tasks, long analyses, permanent archives, external replication.

SUMMARY

What the third official test shows

In four days the Model Lab has moved from a certified measurement to a result announced in advance and proven in public, and LILI's central promise now rests on two independent tests: the accuracy of the best models at a fraction of their cost.

99 / 99
every task solved, level with Claude Opus 5.5 and GPT-5.5 (NEO AUTO)

2.1x
lower cost per successful task than Claude Opus 5.5; 2.9x than GPT-5.5

2 / 2
targets published before the run, both met with room to spare

12 proofs
from four outside parties and a public register, checkable by anyone

Accuracy holds where it matters

NEO AUTO answers 99 of 99 tasks exactly, on reasoning, data and code, at less than half the cost of the leading models and faster than both.

The mechanism is visible

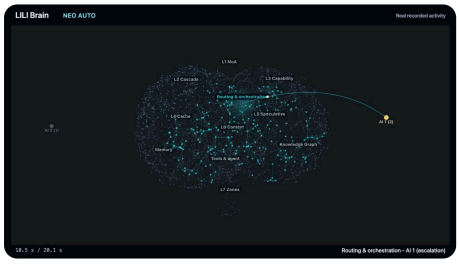
On 66 of 99 tasks LILI verified that a fast-tier answer was right. The saving is the cost of the calls it did not need to make.

Three profiles, one brain

NEO PRIME is the lowest-cost system of the test and NEO PULSE the fastest; NEO AUTO holds top-model accuracy on demanding work.

The promise was made before the result

Targets time-stamped by outside authorities before the run, results sealed when it ended, and a public pack that anyone can recompute.



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, to the millisecond. Those live recordings are replayed at real speed in two films: a NEO AUTO task and a NEO PULSE task of this test. Nothing is staged: the films show what LILI actually did - routing, the first answer, the independent verification and, when the verifier asks for it, the call to the top tier. The AI services appear under neutral names. Watch them on lilineo.com/evaluations/official-tests/test-3/.

This was LILI's third official test. It confirms on a second, harder family of tasks what the second test showed on analysis work: real production data, sealed and certified in time, and now a target announced in advance and met in public. LILI and the Model Lab both continue to improve, and each official test will be reported with the same rigour and the same proof.

ANNOUNCED
SEALED
PROVEN

Important notice. This report presents the results of an internal NEO AI test run on 22 November 2026 on 99 tasks with one known answer and five systems, with each LILI profile as configured on that date. 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; no confidence interval is claimed. The three leading systems each solved every task: the test therefore measures a difference of cost and time at equal accuracy on this set, not a difference of accuracy. The two targets were set after one calibration pass on the same tasks on the same day, and published and time-stamped before the official run. The counts of LILI's decisions are read from the platform's own records and are not part of the figures recomputed by the public verifier. Passages marked as interpretation are opinions. 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 named are independent providers of those public services and have not reviewed this report. 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.