NEOLABS · OFFICIAL LILI CAPABILITY TESTS

Official Test No. 2 – 18 November 2026

34 analysis tasks · 5 systems · 170 completed answers

LILI and two leading single models answered the same 34 analysis tasks through real calls at real prices. Every answer was timed and priced by the platform, scored blind by two independent AI reviewers, sealed in the Proof Ledger and time-stamped by DigiCert in the second the test ended.

4.9×
lower cost per answer for NEO AUTO than Claude Opus 5.5 (0.018 vs 0.090 USD)

12 / 12
cost and speed comparisons statistically significant, all in LILI's favour

+0.95
points out of 10: NEO AUTO judged better than GPT-5.5 (95% interval 0.28 to 1.63; n = 32 paired tasks)

12:10:44 UTC
the sealed results time-stamped by DigiCert, in the second the test ended

RESULTS

Reference-level quality at about one fifth of the cost.

NEO AUTO, LILI's balanced profile, against Claude Opus 5.5, the strongest reference on this test.

NEO AUTO against the strongest reference - 34 tasks; quality on 32 scored tasks per system

MEASURE	NEO AUTO	CLAUDE OPUS 5.5	READING
Cost per answer	0.018 USD	0.090 USD	4.9× lower cost
Time to a complete answer (median)	19.0 s	32.9 s	1.7× shorter time
Judged quality (out of 10)	8.70	8.91	No statistically significant difference detected in judged quality on this test.

Judged quality: No statistically significant difference detected in judged quality on this test. Against GPT-5.5, NEO AUTO was judged better by +0.95 points out of 10 (95% interval 0.28 to 1.63; n = 32 paired tasks; significant after Holm's correction).

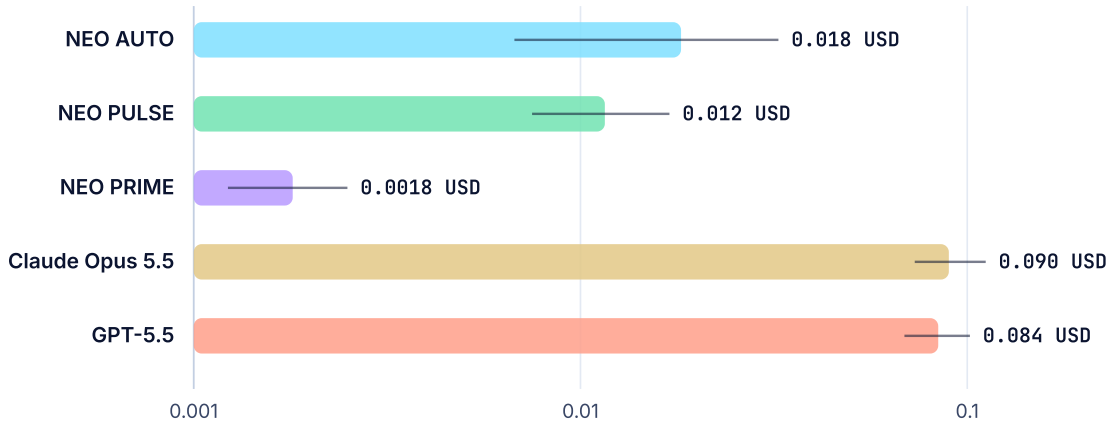
All five systems - quality: mean [95% interval] and scored tasks; cost: average per answer; time: median to a complete answer

SYSTEM	JUDGED QUALITY /10	COST PER ANSWER	MEDIAN TIME	BEST SUITED TO
NEO AUTO	8.70 [8.30-9.03] n=32	0.0182 USD	19.0 s	Demanding work where quality must hold
NEO PULSE	6.26 [5.64-6.86] n=33	0.0116 USD	1.6 s	Instant replies, simple requests
NEO PRIME	5.32 [4.73-5.94] n=33	0.0018 USD	2.8 s	Very high volume at minimal cost
Claude Opus 5.5	8.91 [8.56-9.20] n=32	0.0896 USD	32.9 s	Reference
GPT-5.5	7.82 [7.26-8.35] n=33	0.0841 USD	26.2 s	Reference

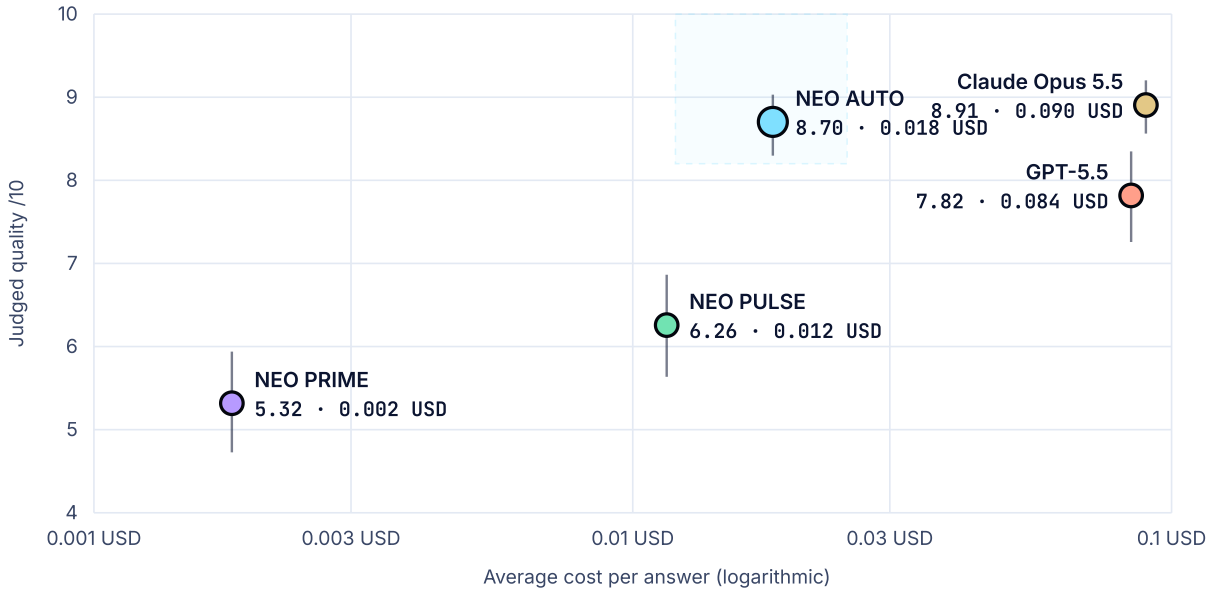
COMPARATOR

Cost, speed and quality, side by side.

Switch the measure. Bars show the value; the white line, its 95% interval.



Average cost per answer, USD - logarithmic scale: each gridline is 10 times the previous one. NEO AUTO: 4.9× lower than Claude Opus 5.5.



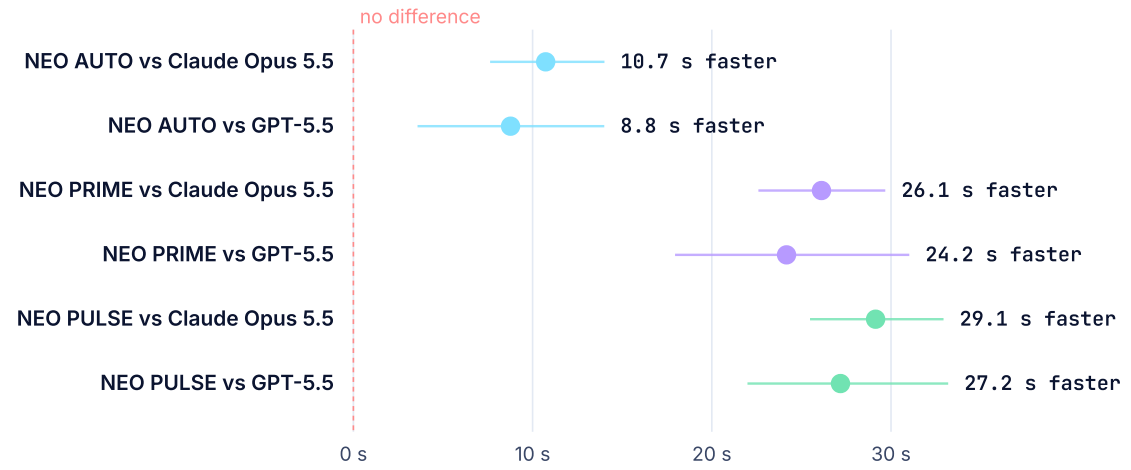
Ratios of the average cost and of the median time, rounded. All twelve underlying differences are statistically significant.

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	4.9×	4.6×	1.7×	1.4×
NEO PULSE	7.8×	7.3×	20×	16×
NEO PRIME	50×	47×	12×	9.5×

STATISTICS

12 of 12 cost and speed comparisons significant, all in LILI's favour.

Every comparison is paired on the same task and corrected for the 24 comparisons made at once. 17 of 24 are significant, 13 of them in LILI's favour.



Dot: mean paired difference over 34 tasks. Bar: 95% interval. A bar that does not touch zero is significant.

The 12 significant cost and speed comparisons (mean paired difference, 95% interval, paired tasks)

COMPARISON	MEASURE	DIFFERENCE	95% INTERVAL	N
NEO AUTO vs Claude Opus 5.5	Total time	10.73 s faster	7.63 to 14.01	34
NEO AUTO vs Claude Opus 5.5	Cost per answer	0.0714 USD cheaper	0.0506 to 0.0967	34
NEO AUTO vs GPT-5.5	Total time	8.77 s faster	3.58 to 14.00	34
NEO AUTO vs GPT-5.5	Cost per answer	0.0659 USD cheaper	0.0488 to 0.0839	34
NEO PRIME vs Claude Opus 5.5	Total time	26.12 s faster	22.60 to 29.68	34
NEO PRIME vs Claude Opus 5.5	Cost per answer	0.0878 USD cheaper	0.0716 to 0.1102	34
NEO PRIME vs GPT-5.5	Total time	24.17 s faster	17.95 to 31.02	34
NEO PRIME vs GPT-5.5	Cost per answer	0.0823 USD cheaper	0.0664 to 0.1001	34
NEO PULSE vs Claude Opus 5.5	Total time	29.14 s faster	25.48 to 32.93	34
NEO PULSE vs Claude Opus 5.5	Cost per answer	0.0781 USD cheaper	0.0624 to 0.0972	34
NEO PULSE vs GPT-5.5	Total time	27.18 s faster	21.99 to 33.19	34
NEO PULSE vs GPT-5.5	Cost per answer	0.0726 USD cheaper	0.0576 to 0.0892	34

Judged quality: the 5 significant comparisons

COMPARISON	DIFFERENCE /10	95% INTERVAL	N
NEO AUTO vs GPT-5.5	+0.95 points (LILI judged better)	0.28 to 1.63	32
NEO PRIME vs Claude Opus 5.5	-3.64 points (reference judged better)	-4.41 to -2.91	32
NEO PRIME vs GPT-5.5	-2.50 points (reference judged better)	-3.12 to -1.89	33
NEO PULSE vs Claude Opus 5.5	-2.72 points (reference judged better)	-3.44 to -1.92	32
NEO PULSE vs GPT-5.5	-1.56 points (reference judged better)	-2.19 to -0.89	32

Seven comparisons are not shown: NEO AUTO against Claude Opus 5.5 on judged quality (no statistically significant difference detected in judged quality on this test) and the six time-to-first-text comparisons, left for a coming test once every profile streams its answers. Under NEOLABS rules a result that is not significant is never reported as a tie or a win. NEO PULSE and NEO PRIME are built for instant and high-volume work; on these demanding analysis tasks the references keep the lead on judged quality, which is why LILI fits the profile to the request.

83%

of judgements returned: 281 of 340. 32 of 34 tasks have every system scored.

76.3%

of the 118 answers scored by both reviewers are within one point (mean difference 1.04; correlation 0.718).

78.1%

within one point on the position check (32 pairs), which guards against order effects.

THE EIGHT NEO LAYERS

4 of 8 layers at work in this test.

LILI is a layer of intelligence above the best models of every provider: it decides how much compute each request deserves, verifies the result, remembers, and proves what it did.

LAYER 1

Dynamic Capability Routing

Reads each request and selects the tier, the tools and the data it needs, before any model is paid.

✓ At work: 34 requests routed: 32 to the fast tier first, 2 directly to the top tier

LAYER 2

Adaptive Model Cascade

A fast, low-cost model answers first. A top-tier model is called only when the answer needs it.

✓ At work: 32 first answers by the fast tier; 2 escalated to the top tier

LAYER 3

Verification-Driven Compute

An independent model from another provider scores the answer. Evidence decides whether more compute is spent.

✓ At work: 32 answers verified and accepted

LAYER 4

Memory and Context Optimization

LILI sends the facts that matter from its memory graph and a running summary, instead of the whole history.

✓ At work: 34 requests answered with graph facts

LAYER 5

Mixture of Agents

For expert work: up to three models from different providers answer in parallel, then one synthesis and one critic.

Built; not at work on this test

LAYER 6

Speculative Parallel Intelligence

Two approaches are sketched in parallel and scored blind; the stronger one is developed.

Built; not at work on this test

LAYER 7

Semantic Compute Reuse

A verified answer to an equivalent question is reused for the same person instead of being computed again.

Built; not at work on this test

LAYER 8

Hybrid Execution

Designed to place models on API, private or self-hosted capacity. Today every model runs through APIs.

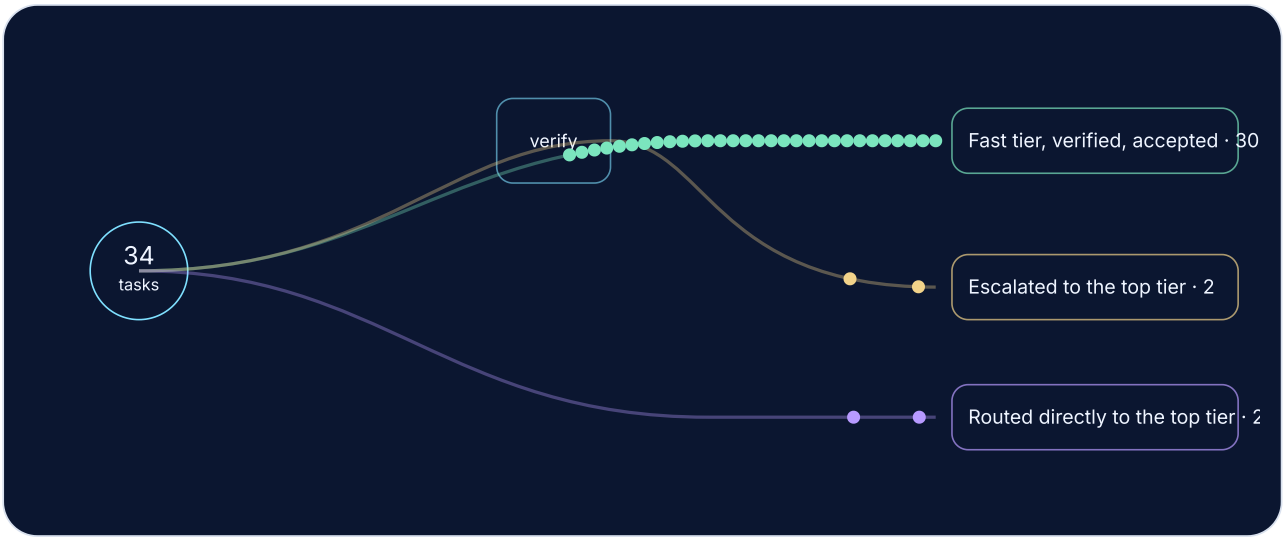
Built; not at work on this test

Layers 5 to 8 are built and used by other profiles and request types; they were not at work on this test.

LILI DECISIONS

The saving comes from the calls LILI did not make.

The saving comes from calls that were not made: on 30 of 34 tasks LILI established by verification that a fast-tier answer was good enough. Every escalation is counted in the averages: the 4.9× advantage is net of them.



What LILI decided, across the 34 tasks (NEO AUTO)

WHAT LILI DID	TASKS	TYPICAL COST	TOTAL
First answer by a fast-tier model, verified and accepted; top-tier call avoided	30	0.0027 USD	0.123 USD
First answer judged insufficient by LILI; escalated to the top-tier model	2	0.141 USD	0.281 USD
Request routed directly to the top-tier model, then verified	2	0.107 USD	0.215 USD
NEO AUTO, 34 answers	34	0.018 average	0.619 USD
Claude Opus 5.5, 34 answers	34	0.090 average	3.047 USD

One answer, opened completely: Task 19

> *A founder wants one assistant that remembers users across five different products. Give the strongest argument for a single shared memory, the strongest argument for keeping each product's memory separate, and the design you would recommend, with the consent and deletion rules it needs.*

1

Request arrives

The question is received with its context. Nothing is sent to a model yet.

`0 USD`

2

Dynamic Capability Routing

LILI classifies the request and selects a fast-tier model for the first answer.

`0 USD - no model call`

3

First answer

The fast-tier model writes the full answer: 2,054 output tokens.

`0.00156 USD`

4

Verification

An independent model from a different provider scores the answer: 1.0 out of 1.0, accepted.

`0.00056 USD`

5

Adaptive Model Cascade

The cascade was armed. Because the answer was accepted, the top-tier model was not called.

`call avoided`

About 30× lower cost than Claude Opus 5.5 on this request, for the same judged score (9.0). Recorded in Proof Ledger entry no. 1143.

SYSTEM	CALLS	TIME	COST	JUDGED /10
NEO AUTO	2	13.6 s	0.00212 USD	9.0
Claude Opus 5.5	1	30.8 s	0.06315 USD	9.0
GPT-5.5	1	34.1 s	0.08114 USD	7.0

INTEGRITY

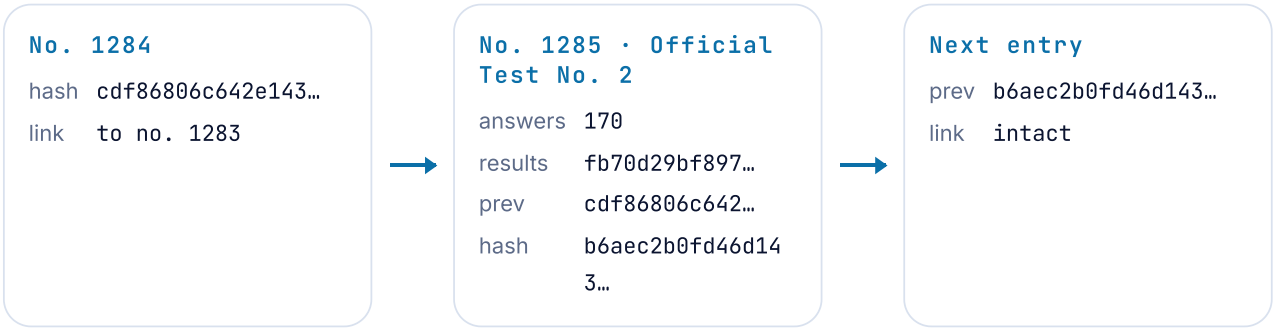
Sealed, chained and time-stamped.

The fingerprint of all 170 answers and 396 judgements was sealed in one ledger entry, then certified in time by DigiCert.

LEVEL 1 · PROOF LEDGER

Sealed in a hash-chained register

Every official test is recorded in NEOLABS' built-in Proof Ledger. Cryptographic hashing makes changes to sealed records detectable, so the integrity of the recorded results can be checked.



This demonstration recomputes the real SHA-256 of entry no. 1285 in your browser. Altering one figure (170 answers becomes 171) changes the hash, and every later link breaks.

LEVEL 2 · INDEPENDENT TIMESTAMP

Certified in time by DigiCert

In the second the campaign ended, the hash of its sealed entry was sent to DigiCert, a public certificate authority, under RFC 3161, the international standard for trusted time-stamps. DigiCert signed the hash with its own clock. Only 32 bytes left NEO AI: no question, no answer, no figure, no name.

The results cannot be changed afterwards. Any change to an answer, a score or a cost changes the fingerprint; the token would no longer match.

The date cannot be moved. No one can obtain a token with an earlier time.

The whole history is covered. Each entry contains the hash of the one before, so one token covers all 1,285 entries up to this test.

Results timestamped by DigiCert · RFC 3161

Official Test No. 2 – proof of record

Time certified	18 November 2026, 12:10:44 UTC
Proof Ledger entry	No. 1285 (lab.campaign, sealed 2026-11-18 12:10:44.240 UTC)
Token serial	`5da7bc65955b06f9c5efc6871bbf4a10`
Hash signed	`b6aec2b0fd46d143fac82b8ba9c63e0edf960084c93b2fbb24456fd2baedb7f2`
Signed by	DigiCert SHA256 RSA4096 Timestamp Responder 2026 1 · RFC 3161 · root DigiCert Trusted Root G4

Third-party timestamp
Verified: DigiCert RFC 3161 token over ledger entry no. 1285.

External audit
Not yet performed. The test has not been audited by an external party.

Independent replication
Not yet performed. The test has not been reproduced by an outside party.

Signature of the token

The cryptographic signature is checked with free software and DigiCert's public root certificate. One command, without NEO AI:

```
openssl ts -verify -in test-2-ledger-1285.tsr -digest b6aec2b0fd46d143fac82b8ba9c63e0edf960084c93b2fbb24456fd2baedb7f2 -CAfile digicert-trusted-root-g4.pem
```

Expected answer: `Verification: OK`. Root certificate SHA-256 fingerprint: `55:2F:7B:DC:F1:A7:AF:9E:6C:E6:72:01:7F:4F:12:AB:F7:72:40:C7:8E:76:1A:C2:03:D1:D9:D2:0A:C8:99:88`.

Proof files

Proof Ledger entry no. 1285 — ledger-entry-1285.json
`SHA-256 62a31f1a47a1cb5bdd7fd1ecb6e2c964262807650b1f39d695fb40e48da5bf7d`

Recompute the entry hash (Node.js) — verify-entry.mjs
`SHA-256 3861bdc3ea704b9d29e731f025864359b13f7c552afc982a4e4a6fe84425c384`

DigiCert RFC 3161 time-stamp token — test-2-ledger-1285.tsr
`SHA-256 e9bd2ca9df8fd9b49af122eb887644e09b15fb6244eaf7dfcaa93eff979366b6`

DigiCert Trusted Root G4 (public root certificate) — digicert-trusted-root-g4.pem
`SHA-256 ce7d6b44f5d510391be98c8d76b18709400a30cd87659bfebe1c6f97ff5181ee`

The results digest (fb70d29b…8c5b) covers the 170 answers and 396 judgements; recomputing it needs those records, which are available on request under confidentiality. The ledger entry and the token can be

checked by anyone with the files above.

TIMELINE

From a frozen task set to a certified record.

2026-11-17

Task set v4 frozen

34 original analysis tasks version-locked before the test.

2026-11-18 10:27:25 UTC

Campaign starts

The first of 170 live calls leaves the Model Lab.

2026-11-18 12:10:44 UTC

Campaign ends

170 answers and 396 judgements stored; their fingerprint is computed.

2026-11-18 12:10:44 UTC

Sealed in the Proof Ledger

Entry no. 1285, chained to the 1,284 entries before it.

2026-11-18 12:10:44 UTC

Time-stamped by DigiCert

RFC 3161 token, serial 5da7bc65955b06f9c5efc6871bbf4a10. Only the 32-byte hash left NEO AI.

2026-11-18

Published by NEOLABS

Methodology, results and proof files made public.

METHOD & LIMITS

What this test shows, and what it does not.

How every figure was produced

No figure was typed by a person. Each one is computed from records written by the platform while the test ran: the frozen task set, live production calls, the tokens reported by each provider, the server's timing events, the blind review and the statistical engine.

A comparison pairs one LILI profile with one reference model on the tasks both answered.

Paired sign-flip test, 20,000 flips, seeded; significant at 0.05 after Holm's correction over all 24 comparisons of the run.

At least 10 paired tasks; the size is the mean paired difference with its 95% bootstrap interval.

Anything else is reported as not shown, never as a tie or a win.

Limits

One run, on one day, on the network of the server: times include the providers' queues and change from day to day.

Quality is judged by AI models, not by people: two reviewers from other providers than the answers, blind to the system; their agreement is reported.

Costs are priced from the tokens reported by each provider and the price table in force on the day.

Some providers do not stream their answers: time to first text is not compared in this test.

Each LILI profile is measured as configured on 18 November 2026; another configuration needs another run.

The time-stamp certifies that the recorded results existed, unchanged, at the stated time. DigiCert provides that public service and has 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.

Claude Opus 5.5 and GPT-5.5 are products of their respective owners, used as references through their public interfaces.

Next milestones

Breadth

More task families after 34 analytical tasks: reasoning, code, long documents and more languages.

Review

Two independent AI reviewers score blind today; human experts will be added on a sample.

Precision

Larger samples, to measure the gap with the strongest reference more finely.

Independence

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