LINUR

Scientific benchmark public report

Chart-led edition: science, evidence and performance at a glance

Headline result. LINUR, Claude Opus 5 and GPT-5.6 Sol each passed all 77 scored fields across five source-traceable scenarios.

Protocol qualification pilot / 15 preserved runs / PQP5-3.0.0-rc.4 / Report v6

1. Executive summary

The four charts below summarize timing, correctness, accuracy and numerical precision. Colors identify systems only; they do not encode rank or preference.

Timing

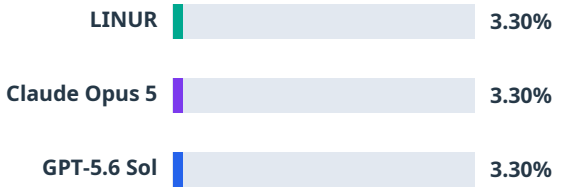
Log bars from 0.01 s / source-labelled

Correctness

Scored fields passed

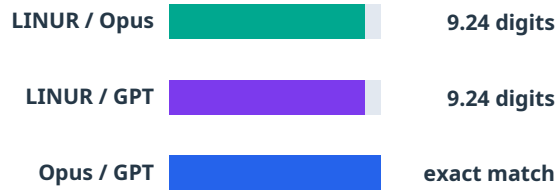
Accuracy

Largest error as % of allowed tolerance



Precision proxy

Equivalent digits: -log10(max relative difference)



Main finding. All systems completed every case and passed every scored field. The scientific result is a tie. Timing differs strongly, while the available compute measures use different units and scopes.

Timing and precision boundary. LINUR timing covers its local process; Opus/GPT timing covers the full model session including orchestration. The timing chart is not hardware-normalized. Precision shows cross-system numerical agreement; repeatability is not estimated from one run per cell.

2. Topics tested and reference basis

All five topics are shown on one page. Input references identify the data supplied to respondents; correctness references identify the sources and deterministic procedures used to construct the sealed referee values.

Case	Field / bottleneck	Problem	Input reference	Correctness reference
S004	**Materials science** Nonlinear EOS fitting	Fit six energy-volume curves; aggregate crystal groups and deltas.	[I1] AiiDA/GPAW unary + oxide data; eosfit 3.1.	[C1] Deterministic recomputation from the same pinned data and fit code.
S007	**Molecular simulation** PMF feature extraction	Extract basins, barriers, trends and artifact concerns at two temperatures.	[I2] DTU TIP4P/2005 PMF dataset.	[C2] Deterministic extraction from pinned Figshare files.
S012A	**Quantum transport** Forty-array reduction	Measure the conductance effect of twist-angle disorder in TBG.	[I3] Zenodo 13972801 arrays and ensemble definitions.	[C3] Reductions using the deposited notebook convention.
S019B	**Nuclear modelling** Coupled-chain propagation	Propagate xenon/cesium depletion with continuous xenon removal.	[I4] OpenMC 0.15.2 depletion and transfer inputs.	[C4] Deterministic solution from pinned OpenMC regression artifacts.
S042	**Micromagnetics** Unit-safe field calculations	Compute anisotropy fields and exchange lengths for five compounds.	[I5] Pinned magnetic_dataset material descriptors.	[C5] Deterministic evaluation of declared SI equations.

Citations

[I1] AiiDA ACWF GPAW unary/oxide PBE data and eosfit 3.1, commit 4e26944392a13841289d4f31e813b1dd665ecf4b.

[C1] Same pinned AiiDA/GPAW artifacts and fitting implementation; context: Bosoni et al. (2024), doi:10.1038/s42254-023-00655-3.

[I2] Muthachikavil et al., PMF Files (TIP4P/2005), DTU (2021), doi:10.11583/DTU.14138057.v1.

[C2] Deterministic feature extraction from pinned Figshare files 26659037 and 26659043 under doi:10.11583/DTU.14138057.v1.

[I3] Sanjuan Ciepielewski et al., Transport effects of twist-angle disorder in mesoscopic TBG, Zenodo (2024), doi:10.5281/zenodo.13972801.

[C3] Deterministic reduction of the same deposited arrays using the published notebook convention, doi:10.5281/zenodo.13972801.

[I4] OpenMC 0.15.2 source-native depletion and transfer regression inputs, Git commit e23760b0264c66fb7bb373aa0596801e5209d920.

[C4] Deterministic coupled-chain solution from the same pinned OpenMC source inputs and regression definitions.

[I5] Bhandari and Nop, magnetic_dataset, Git commit a7db2e17e016dab609251085ebc6dde6c129aa05.

[C5] Deterministic evaluation of the declared SI equations using the same pinned magnetic-material descriptors.

3. How the test works

Each respondent receives the same frozen public package. Responses are structurally validated before a private scorer compares every declared output with a sealed, source-derived reference and predeclared tolerance.

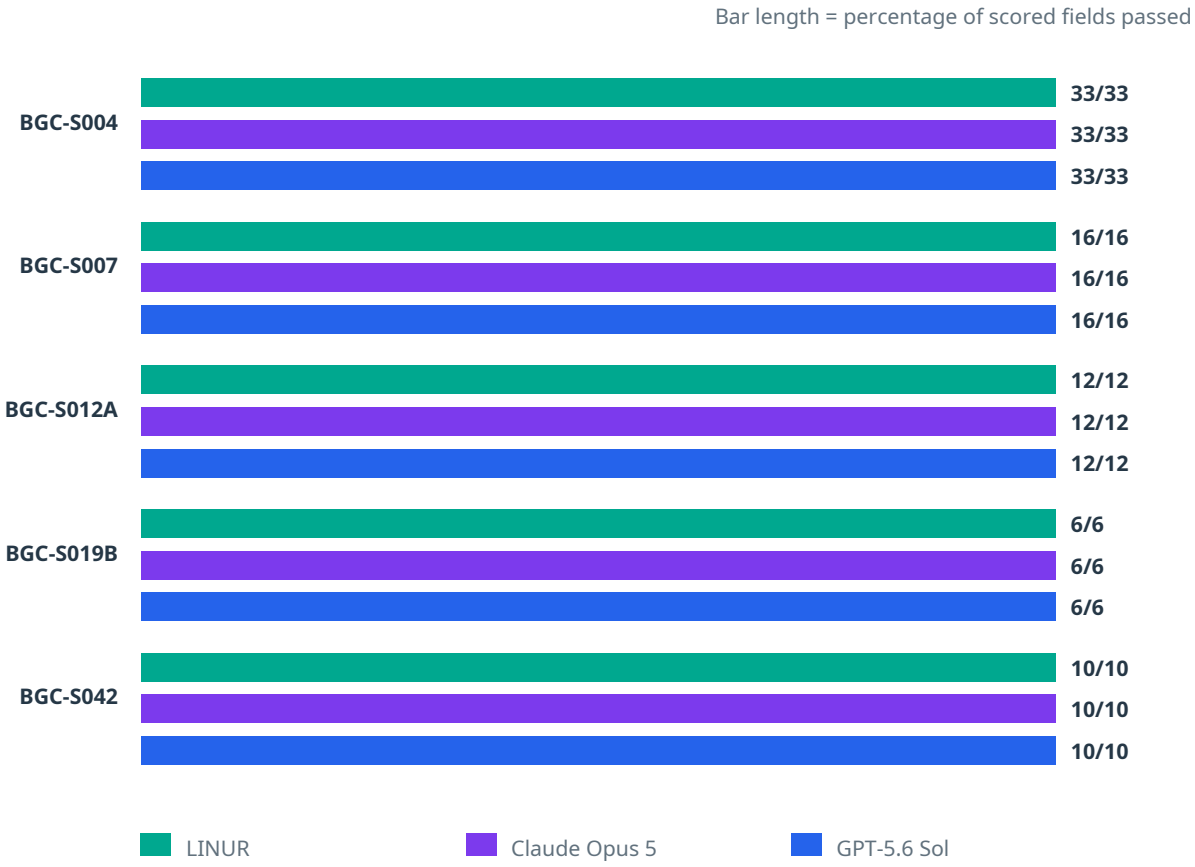
Step	Control	Why it matters
1	Pin source artifacts	Prevents source drift
2	Freeze prompt and inputs	Keeps tasks identical
3	Run in isolation	Prevents answer sharing
4	Validate schema and hashes	Rejects malformed evidence
5	Score against sealed references	Measures correctness
6	Publish aggregate evidence	Protects reusable answers

Correctness Pass/fail against the declared scientific acceptance rule.	**Accuracy** Distance from the sealed reference, normalized by the allowed tolerance.
Precision Agreement of reported numerical values across systems. Repeatability requires more runs.	**Performance** Reported elapsed time and native compute/resource evidence, with source and scope retained.

Scientific interpretation. The references are deterministic calculations or extractions from pinned scientific sources. Passing demonstrates reproduction of the benchmark result, not a new experimental validation of the underlying physics.

4. Scientific results

Every model is scored independently. Grouped bars show correctness; the table shows pairwise numerical agreement.

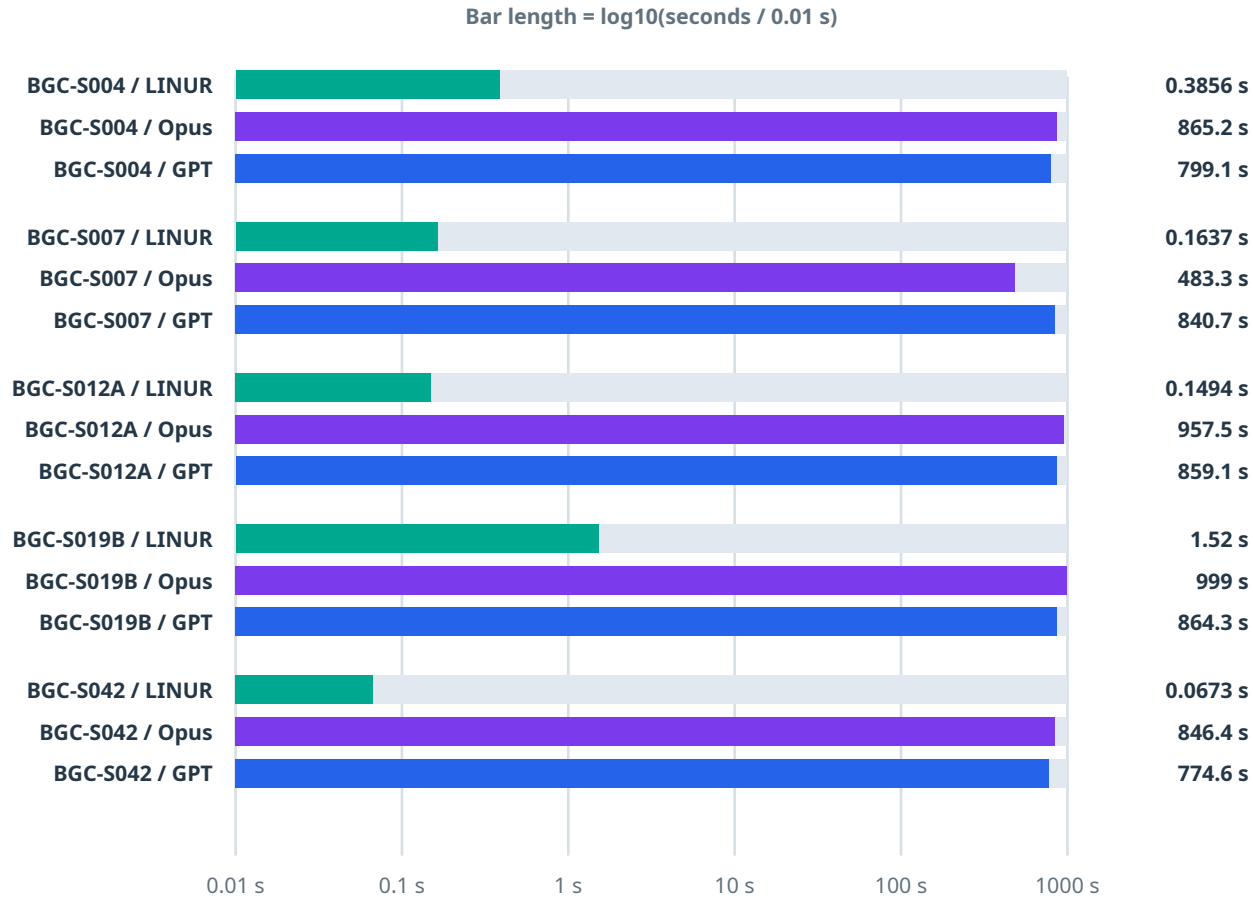


System pair	Numerical fields	Largest relative difference
LINUR / Claude Opus 5	73	5.72e-10
LINUR / GPT-5.6 Sol	73	5.72e-10
Claude Opus 5 / GPT-5.6 Sol	73	0

Result. All 231 scored respondent-field instances passed. All 12 pairwise categorical comparisons matched. Three artifact-concern comparisons are reporting-only.

5. Reported elapsed time

The combined log-scale chart makes the reported differences visible while retaining the evidence source for every series.



Timing boundary. LINUR reports local dispatch to local output. Opus and GPT use full session dispatch to complete output receipt, including orchestration. The chart is an operational comparison, not a hardware-normalized speed claim.

Scenario	LINUR	Claude Opus 5	GPT-5.6 Sol
BGC-S004	0.3856 s	865.2 s	799.1 s
BGC-S007	0.1637 s	483.3 s	840.7 s
BGC-S012A	0.1494 s	957.5 s	859.1 s
BGC-S019B	1.52 s	999 s	864.3 s
BGC-S042	0.0673 s	846.4 s	774.6 s

6. Compute and resource evidence

Hosted models and the local LINUR process expose different native measurements. These values are shown directly and are not forced into one synthetic efficiency score.

Hosted-model provider accounting

Model	Provider time	API calls	Input tokens	Output tokens	Cache-read	Nano AIU
Claude Opus 5	1714.4 s	84	6,139,890	79,790	5,856,737	6.693e+11
GPT-5.6 Sol	916.6 s	39	1,757,271	37,693	1,295,872	3.118e+11

Provider time sums model-call durations and excludes queueing and orchestration. Nano AIU is an internal accounting unit, not FLOPs or energy.

LINUR local diagnostics

Scenario	CPU time	Peak RSS	Energy	Power	GPU time
BGC-S004	0.103 s	34.5 MB	N/A	N/A	N/A
BGC-S007	0.08253 s	25.1 MB	N/A	N/A	N/A
BGC-S012A	0.09279 s	34.9 MB	N/A	N/A	N/A
BGC-S019B	0.4312 s	78.5 MB	N/A	N/A	N/A
BGC-S042	0.03253 s	21.2 MB	N/A	N/A	N/A

Energy status. No respondent supplied measured energy, average power or GPU time. Those fields remain N/A; no value is estimated and no energy-efficiency claim is made.

7. Conclusions and publication boundary

Supported conclusions

- All three systems reproduced every scored canonical result.
- Numerical agreement is extremely close in this five-case sample.
- Reported timing and provider-native compute evidence are documented.
- Every source used for inputs and correctness is publicly cited.

Not established

- Portfolio-wide superiority from five selected scenarios.
- Repeatability precision from one accepted run per cell.
- Hardware-normalized speed across different execution stacks.
- Energy efficiency without measured energy.
- Independent experimental validation of the underlying physics.

Bottom line. Within this protocol-qualification pilot, all three systems were scientifically equivalent under the declared rules. Operational differences are visible, but their evidence sources and execution boundaries must travel with the values.

Public release contents

Item	Publication status
Aggregate scientific scores	Included
Input and correctness citations	Included
Source-labelled timing and compute evidence	Included
Sealed numerical references	Excluded
Response-bearing private comparison JSON	Excluded