LEAP-BENCH: TURNING RAW DATA INTO SCIENTIFIC CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific discovery depends on turning raw experimental data into defensible claims through quantitative analysis and physical interpretation. Whether AI agents can carry out this process reliably remains unclear. We introduce LEAP-Bench, a benchmark of end-to-end experimental analysis in materials science. We construct 64 tasks from 12 papers spanning multiple instrument modalities, each task targeting a central scientific claim. Agents analyze raw experimental files without seeing the papers' conclusions or procedures. Using reference answers from the source papers, 475 questions assess whether agents extract the required quantities, correctly reason through the analysis and justify their scientific conclusions. Across three evaluation rounds, Claude Code with Opus 5 achieves the highest task success rate among the six evaluated agents, at $59.4\% \pm 7.2\%$ (mean $\pm$ standard deviation). Even when agents correctly answer questions about numerical results and intermediate reasoning, their scientific interpretations can remain incomplete. Twelve tasks remain unsolved by all six agents across all three rounds, including eight requiring model fitting or comparison across measurements. These findings highlight persistent difficulties in completing experimental analyses and connecting the results to scientific conclusions. By tracing reasoning from data to claims, LEAP-Bench provides a basis for training AI co-scientists through feedback on numerical estimates, analytical reasoning and physical interpretation.[1]

1 INTRODUCTION

Interpreting experimental measurements is central to autonomous materials research because analytical conclusions guide both the assessment of an experiment and the choice of subsequent experiments (Szymanski et al., 2023). Systems that combine language models, analysis tools and laboratory interfaces have already planned and executed chemical syntheses, illustrating how AI agents can connect scientific reasoning to experimental work (Boiko et al., 2023; Bran et al., 2024).

Coding agents offer a practical interface for scientific analysis through file access, program execution and iterative refinement (Chen et al., 2025). Applications in the life sciences and chemistry also use literature search, databases and analysis tools (Anthropic, 2026; Mitchener et al., 2025; Boiko et al., 2023; Bran et al., 2024), supported by shared data archives and standard representations (Berman et al., 2000; Benson et al., 2013; Heller et al., 2015; Kim et al., 2023). These applications motivate testing whether agents can select and execute an appropriate analysis of unfamiliar experimental data.

Existing materials-science systems typically target a particular technique or research stage, including diffraction analysis (Wu & Sun, 2026), spectroscopy (Xie et al., 2026), and instrument control (Zhu et al., 2024; Pratiush et al., 2025). Autonomous laboratories connect synthesis and testing through analysis pipelines tailored to their platforms (Szymanski et al., 2023; Zhang et al., 2025; Song et al., 2025). Analysis across laboratories must accommodate different file formats, measurement conventions and sources of experimental artefacts. These differences make the choice and interpretation of an analysis central to evaluating scientific agents.

We study end-to-end reasoning from experimental data to scientific claims. An agent must inspect measurement files, including raw instrument data, choose an appropriate analysis and explain what the resulting quantities establish about the material. Evaluating only the extracted values misses the

[1]The benchmark, construction pipeline, and per-system evaluation harness will be released upon publication.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

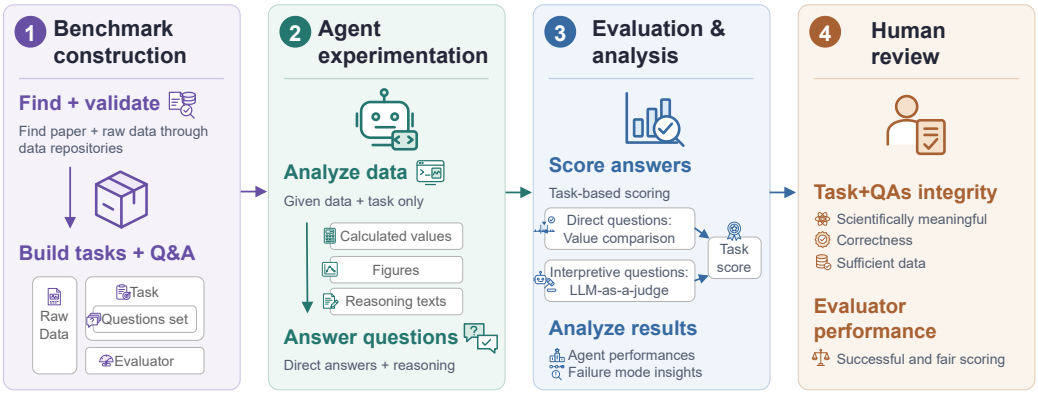


Figure 1: **Overview of LEAP-Bench.** (1) We identify candidate papers and datasets through automated searches, then build analysis tasks and reference questions from published claims and raw measurements. (2) Agents analyze the data before receiving the questions. (3) Numerical checks and an independent language-model judge assess the answers. (4) Expert review examines selected tasks, reference answers and saved analyses.

reasoning that connects those values to a scientific conclusion. Existing benchmarks assess scientific knowledge, published figures, scientific programs or computational reproduction (Section 2). Our focus is the complete experimental analysis across materials-science measurement modalities, with explicit checks along the reasoning from data to claim.

We construct LEAP-Bench around published scientific claims and the measurements needed to support them. Each source paper provides the evaluator with the context, supporting measurements and reference conclusion. The agent receives a research objective and measurement files, with the analytical choices left open. It first carries out the analysis and records its calculations, figures and conclusions; only then does it receive questions about key quantities, reasoning steps and physical interpretation. The resulting feedback identifies gaps between analytical results and scientific interpretation, and could support future training environments for AI co-scientists.

2 Related Work

Knowledge and visual interpretation. MaScQA (Zaki et al., 2024), MatSciBench (Zhang et al., 2026) and OmniMatBench (Liu et al., 2026) evaluate materials knowledge through exam, textbook and expert-curated problems. MaCBench (Alampara et al., 2025), MatVQA (Wu et al., 2026) and MatQnA (Weng et al., 2025) extend evaluation to scientific images and characterization figures across several instrument modalities. These benchmarks assess responses to formulated questions. Our tasks require agents to select and execute an analysis of experimental measurement files, then interpret the results to answer the research question.

Scientific workflows. ScienceAgentBench (Chen et al., 2025) evaluates scientific programs on diverse data, including measured signals, with expert-defined milestone rubrics. BLADE (Gu et al., 2024) grades analytical decisions for research questions, including variable selection, transformations and statistical models. XRDBench (Wu & Sun, 2026) provides measured X-ray and neutron diffraction patterns and grades analysis workflows within powder diffraction. Reproduction benchmarks evaluate implementation and execution with varying access to existing code (Starace et al., 2025; Siegel et al., 2025; Yan et al., 2025; Qiu et al., 2026); PaperBench and PRBench also grade implementation or methodological correctness. AstaBench (Bragg et al., 2026) combines literature, coding and research tasks in a heterogeneous suite, while DiscoveryWorld (Jansen et al., 2024) evaluates scientific discovery in a simulated environment.

Claim discovery. DiscoveryBench (Majumder et al., 2025) evaluates hypothesis discovery from data. FIRE-Bench (Wang et al., 2026) asks agents to rediscover empirical findings while withholding the published conclusions. Their hypothesis and claim scores assess the final scientific findings. LEAP-Bench couples claim recovery with intermediate-answer checks on experimental data.

Table 1: **Comparison with prior benchmarks.** *Raw experiment data*: instrument recordings before scientific reduction. *Multi-instrument*: coverage across measurement families; variants within one family receive partial coverage. *Claim discovery*: infer a scientific finding from data with the target conclusion withheld; selecting among supplied candidate conclusions receives partial coverage. *Scientific interpretation*: assess the scientific meaning or evidential implications of observations and results. *Expert verified*: domain experts check reference analyses or answers, including sampled review. ● documented coverage; ◐ partial coverage; × absent or unreported in the inspected work.

Benchmark	Raw experiment data	Multi-instrument	Claim discovery	Scientific interpretation	Expert verified
ScienceAgentBench (Chen et al., 2025)	●	●	×	×	●
BLADE (Gu et al., 2024)	×	×	×	×	●
DiscoveryBench (Majumder et al., 2025)	×	×	●	●	×
PaperBench (Starace et al., 2025)	×	×	×	×	●
AstaBench (Bragg et al., 2026)	×	×	●	●	●
FIRE-Bench (Wang et al., 2026)	×	×	●	●	●
PRBench (Qiu et al., 2026)	×	×	×	×	●
Materials-science-specific					
MaScQA (Zaki et al., 2024)	×	×	×	×	×
MatSciBench (Zhang et al., 2026)	×	×	×	×	●
OmniMatBench (Liu et al., 2026)	×	●	×	●	●
MaCBench (Alampara et al., 2025)	×	●	◐	●	●
MatVQA (Wu et al., 2026)	×	●	◐	●	●
MatQnA (Weng et al., 2025)	×	●	◐	●	●
AutoMat (Huang et al., 2026)	×	×	×	●	●
XRDBench (Wu & Sun, 2026)	●	◐	◐	●	×
LEAP-Bench (ours)	●	●	●	●	●

Materials analysis and claim reproduction. AutoMat (Huang et al., 2026) evaluates computational claim reproduction. Agents receive the target claim, source paper and optional author artifacts; expert-written reproduction procedures remain hidden. Its evaluation assesses methodological fidelity and scientific rigor through execution traces, artifacts and reports, including interpretation of existing computational outputs. Across multiple experimental modalities, LEAP-Bench supplies a research objective and experimental measurement files, with the source paper and reference conclusion withheld. Agents must choose the analysis and infer the finding supported by the measurements. We check derived quantities, reasoning-step answers and claim interpretation (Table 1).

3 Building LEAP-Bench for End-to-End Claim Discovery

An experimental scientist turns measurements into a claim by deciding which features matter, choosing analyses that test a physical explanation, and judging what the resulting evidence supports. We build LEAP-Bench around this connected process, aiming to develop AI co-scientists that can reason through the experimental basis of a scientific finding. The agent receives raw measurements and a research question, with responsibility for the analysis and interpretation. Evaluation follows the evidence from derived quantities through reasoning steps to the final claim. This creates a basis for training agents whose scientific conclusions are supported by the analyses they perform.

3.1 Problem definition

We ask whether an agent can answer an experimental research question by choosing and executing its own analysis of measurement data. The evaluation must connect the accuracy of that analysis to the scientific meaning of its results. Accurate numerical results can still leave the research question unanswered; a stated conclusion alone does not show which evidence the agent obtained. We therefore evaluate the quantities, reasoning steps and claim interpretation together.

We use a scientific claim as the unit of evaluation because its evidence can span several measurements, comparisons and figures. A published study supplies a reference conclusion and its supporting evidence. We recast that conclusion as an open research question and give the agent the measurement

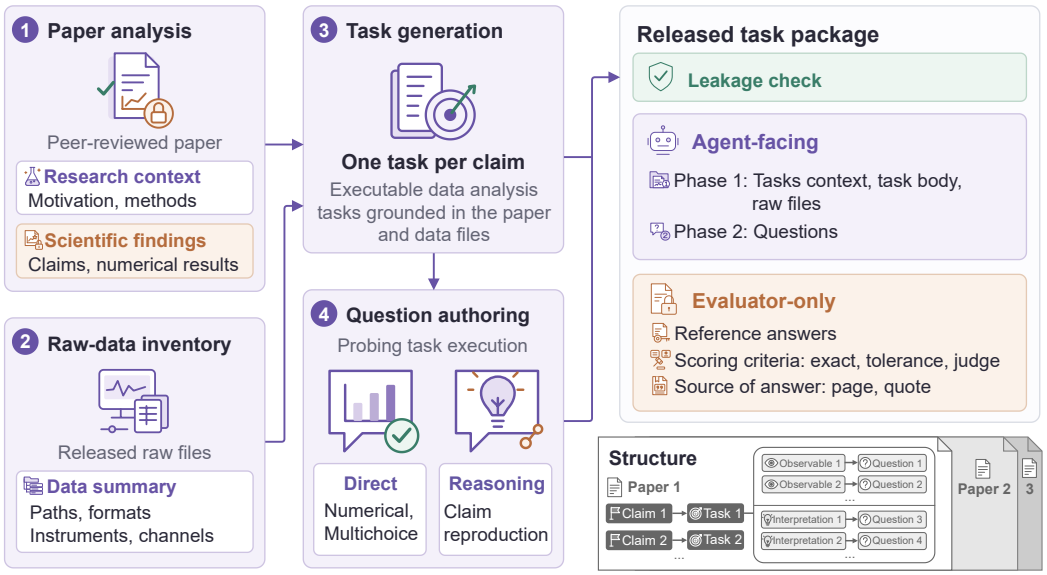


Figure 2: **The LEAP-Bench construction pipeline.** Paper analysis and data inventories link scientific claims to tasks and reference questions, organized in a paper-task-question hierarchy.

files, including raw instrument data, needed to investigate it. The source paper, target conclusion and reference analysis remain with the evaluator.

3.2 Benchmark Construction

The central design challenge is to assess a scientific conclusion while leaving the analysis and interpretation to the agent. We work backward from a published finding to identify its supporting evidence, then define a research question that requires the agent to analyze that evidence. Four stages make these decisions explicit (Figure 2), illustrated by the electrical-measurement task in Figure 3. Language-model agents construct the tasks, and separate model sessions review the experimental inputs, reference answers and task wording. Appendix M documents construction and review.

Paper and data selection. Evaluating reasoning from raw measurements requires access to the experimental inputs behind a published claim. Source-data policies at leading journals encourage authors to share the data underlying their findings,[2] but identifying which released files contain raw measurements requires further inspection. We use a Claude Code pipeline to search Zenodo datasets, link them to papers through OpenAlex, and retain relevant peer-reviewed experimental studies with more than ten citations. The pipeline downloads the associated data and uses paper descriptions and measurement provenance to identify candidate raw files. We manually inspect these files to verify their raw status before task construction.

Stage 1. Paper understanding. The scientific argument determines which measurements belong in a task. We trace a claim to the observations and comparisons supporting it, then identify the corresponding files in the authors' released experimental data. The mapping retains sample identity, measurement conditions and scientific conventions needed to interpret the files in their experimental context. In Figure 3, voltage recordings at opposite magnetic fields and known temperatures provide evidence about the balance of charge carriers in the crystal. This mapping identifies what an agent can investigate from the available measurements and which experimental information it needs.

Stage 2. Task formulation. The task must supply the experimental context needed to reason about the data while leaving the analysis to the agent. We express the selected claim as an open research question. Figure 3 asks how magnetic field and temperature affect carrier balance, leaving the choice of signal processing and physical analysis to the agent. Information derivable from the measurements remains part of that analysis. Necessary factual annotations can be reconstructed and supplied when they do not establish the target conclusion. We narrow or exclude objectives unsupported by the

[2] See the Nature Research data-availability policy.

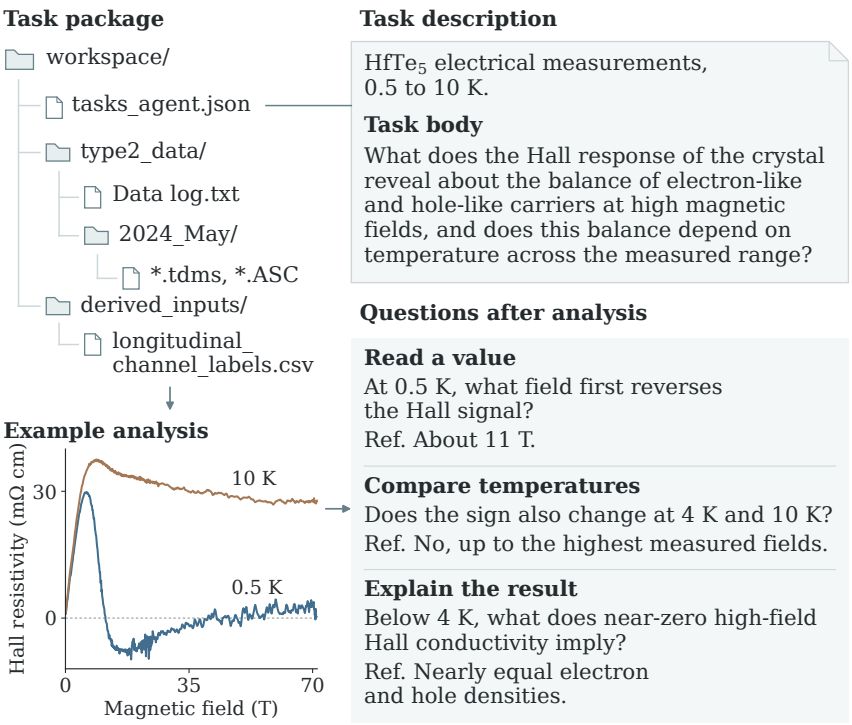


Figure 3: **A task using electrical measurements of $HfTe_5$.** The agent receives measurement files and a research objective, then answers questions about quantities, reasoning and interpretation. The reference plot reconstructs Hall resistivity (transverse electrical response). Electrons and holes carry negative and positive charge. *Ref.* denotes evaluator-only answers. The task body is verbatim; context and questions are abbreviated. Appendix K gives sources and reconstruction details.

available evidence. A separate reviewer checks the required inputs and reconstructs supplementary annotations before consulting their provenance.

Stage 3. Evaluation design. The evaluation checks whether the agent obtains the evidence needed to answer the research question. We author questions about derived quantities, reasoning steps and physical interpretation, with the computational procedure left open. In Figure 3, the field at a sign reversal provides a quantitative result, the comparison across temperatures tests its dependence on conditions, and the interpretation question connects the derived Hall conductivity to carrier balance. These questions connect numerical results to the conclusion they support. Questions specify the sample, conditions and conventions, while reference tolerances reflect reported uncertainty, printed precision or figure resolution. A separate verifier reconstructs answers from the paper before inspecting the proposed references, then confirms, corrects or removes them.

Stage 4. Benchmark assembly. The released task must preserve the agent's responsibility for choosing an analysis. We separate agent inputs from evaluation references and reduce wording that names quantities to compute, prescribes procedures or presupposes the finding. The brief retains the research objective, physical setting and available inputs. A fresh reviewer plans an analysis from the brief and file inventory before seeing the questions, then checks whether the requested evidence follows naturally from that objective. In Figure 3, the research goal and voltage records are supplied for analysis; questions follow afterward, and reference answers remain with the evaluator. This separation makes scientific choices part of the evaluated work.

Expert feedback. Expert review informs the construction requirements and provides a check on the scientific judgments recorded by the pipeline. Reviewer worksheets bring together the target claim, evidence, reference values, tolerance justifications and source passages. Feedback on an earlier release informed the question-purpose definitions, scientific conventions, answer verification and concise task wording. Current expert review examines selected tasks, questions and saved analyses, informing the qualitative findings in Section 4.5.

3.3 TASK INTERFACE AND EVALUATION

The evaluation protocol preserves the agent's responsibility for deciding which analyses the research objective requires. During the analysis phase, the agent receives the brief and measurements, then records its calculations, figures and conclusions. Evaluation questions are revealed afterward and answered from these saved results. This ordering makes the questions checks on the completed analysis without supplying an advance list of quantities to compute. Task type is visible to the agent; reference answers, the reference analysis chain and complexity metadata remain with the evaluator.

We score each answer against the reference and criteria for its question. Closed-form answers use the stored exact-match or numerical-tolerance rule. A separate language-model judge assesses free-text answers against the reference facts, allowing equivalent statements of the same physical meaning. The judge receives the question, reference answer, scoring criteria and agent response, with tools disabled. Its fact ledger records whether required facts are present, equivalent, missing or contradicted, producing a verdict of *pass*, *partial* or *fail*. Explicitly unknown answers receive a fail, and the records are retained for scoring review.

Task completion combines the quantitative and interpretive requirements of the scientific objective. For task t, let R_t contain the reasoning-step and claim-interpretation questions, K_t the key-quantity questions, and $a_q = 1$ denote a pass on question q. With a quantity-error allowance b, task success is

$$S_t = \mathbf{1}[\forall q \in R_t,\ a_q = 1]\,\mathbf{1}\left[\sum_{q \in K_t} (1 - a_q) \le b\right]. \tag{1}$$

For a benchmark with N tasks, the success rate is $100 \sum_{t=1}^{N} S_t / N$. Only pass verdicts count as correct. The setup specifies b; Appendix I gives the scoring diagnostics.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark data and tasks. We evaluate 64 tasks and 475 questions from 12 papers in experimental condensed-matter physics, materials chemistry and electrochemistry. Inputs include raw measurements, instrument exports and selected processed or supplementary files. Tasks require extracting quantities, fitting models, identifying patterns and comparing measurements. Appendices A and M give the corpus composition and input provenance.

Agent configurations. We compare Claude Code with Haiku 4.5, Sonnet 5 and Opus 5 at default reasoning effort, and Codex CLI with GPT-5.6 Luna, Terra and Sol at effort `xhigh`. Both harnesses support file and image inspection and code execution. Claude Code permits subagents of the evaluated model; Codex subagents are disabled. Appendix L records versions, settings and usage accounting.

Execution environment. Runs use isolated workspaces with read-only experimental inputs and writable output space. Source papers and evaluator references are withheld, and web access is disabled. Configurations receive the same inputs and phase instructions within each condition, starting without prior user or project settings or memory.

Evaluation. The primary evaluation uses simplified task descriptions with less guidance on the analysis. We follow the two-phase protocol in Section 3.3, with one persistent session per paper and round. All configurations share the reference answers and tolerances; Claude Fable 5 judges free-text answers with tools disabled. Task success follows Equation 1 with $b = 1$; only pass verdicts count as correct, and tasks receive equal weight. Each configuration is evaluated in three rounds on identical task and question packages. We report the mean and sample standard deviation of the round-level success rates. Appendices I and E give scoring diagnostics and round-specific execution conditions.

4.2 MAIN RESULTS

The strongest agent leaves two in five tasks unsolved. Opus 5 leads in every round, yet completes fewer than 60% of tasks on average (Figure 4A). The other agents solve fewer than half of all tasks. These tasks ask agents to recover established experimental findings while choosing how to analyze the supplied measurements. Their completion rates show that this reconstruction remains difficult under the benchmark's evaluation conditions.

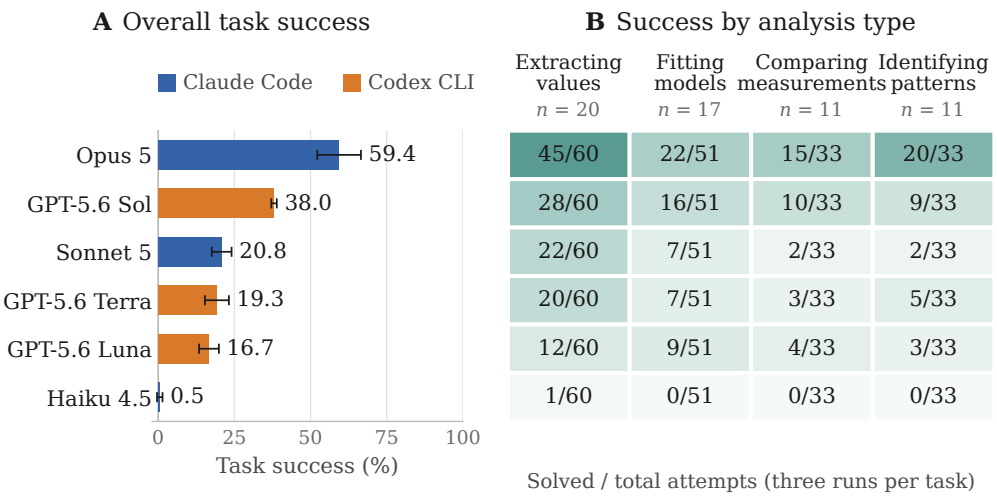


Figure 4: **Overall task success and performance by analysis type.** Model fitting and measurement comparison remain difficult even for the strongest agent. **A** Mean task success over three rounds on 64 tasks; error bars show ± 1 sample SD. **B** Solved / total attempts across the same rounds; n counts distinct tasks. These four analysis types cover 59 tasks. Appendix C defines the categories; Table 5 reports all 64 tasks. Both panels use Equation 1.

Fitting models and comparing measurements remain difficult. Opus solves fewer than half the tasks in model fitting and measurement comparison on average (Figure 4B). Model fitting estimates parameters or scaling laws; measurement comparison reconciles evidence across techniques, samples or conditions. Both require relating observations, beyond identifying individual features. Its higher success in value extraction and pattern identification does not extend to these analyses.

Similar total scores conceal different analytical strengths. Sonnet 5 and Terra have similar mean task success, but Sonnet solves more value-extraction tasks while Terra solves more pattern-identification tasks. Luna exceeds both in model fitting despite its lower overall success. Figure 4B makes these differences visible alongside the overall ranking. An agent's total score therefore does not fully describe which scientific analyses it can complete.

Twelve tasks remain unsolved across models and runs. Repeated execution and changes of configuration leave twelve tasks unsolved across all eighteen combinations of model and round. These tasks span eight papers, including four model-fitting and four measurement-comparison tasks. Their persistence across models and runs supports the difficulty of these analyses beyond an isolated unsuccessful execution. Appendix C reports the complete shared-failure breakdown, and Appendices G and H separate outcomes by paper and reference-chain length.

Passing analysis checks does not ensure correct interpretation. Electrical measurements illustrate a gap between analyzing a signal and explaining what it means for the material (Figure 5). These experiments record current, voltage or resistance. For Sol, interpretation failures when the quantity and reasoning criteria are satisfied are more common than failures of those criteria. For Opus, spectral measurements, which record signals across energy, wavelength or resonance frequency, produce more quantity or reasoning failures than electrical measurements. These patterns separate difficulties in recovering evidence from difficulties in drawing a supported conclusion. Appendix D reports score distributions for each measurement family and explains their relationship to these failure categories.

4.3 Run stability

Stable scores do not ensure repeatable task completion. Sol completes 24, 25 and 24 tasks across the three rounds, yet only 13 tasks succeed in all three. Of the 36 tasks it solves at least once, many do not consistently satisfy the evaluation criteria. This distinction matters when an agent is used to repeat a scientific analysis. A nearly unchanged benchmark score can coexist with failure on tasks the agent previously completed. Table 8 reports the task counts underlying this comparison.

The leading agents stay ahead as nearby ranks change. Opus 5 and Sol rank first and second in every round, while the ordering among Sonnet 5, Terra and Luna changes. Opus also shows the

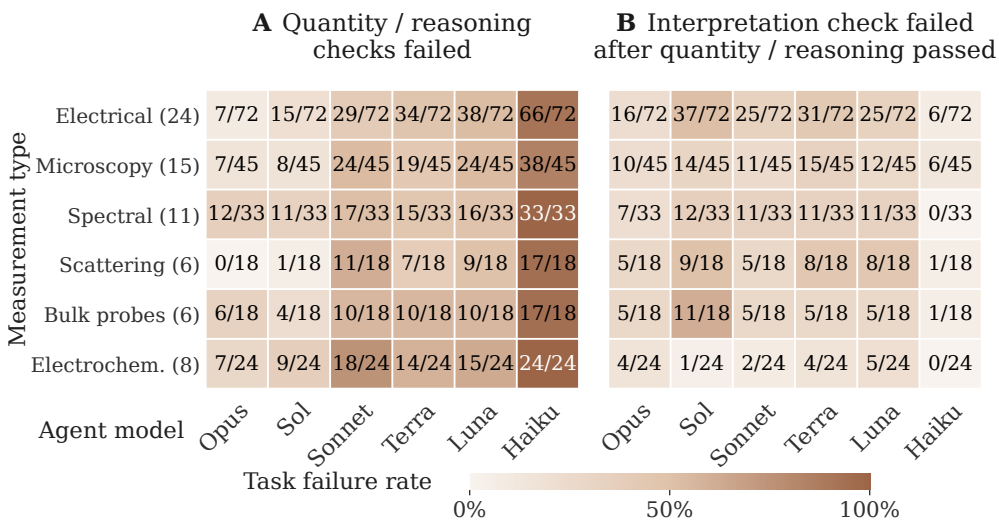


Figure 5: **Failure location across measurement types.** Passing quantity and reasoning checks can still leave interpretation incomplete. Cells give failed / total attempts, using all three runs of every task in the measurement group as the denominator. **A** counts attempts failing quantity or reasoning criteria; **B** counts those passing these criteria but failing interpretation. Parentheses count distinct tasks; tasks using several techniques appear in each applicable group. Task success permits one quantity error under Equation 1.

largest variation in total task success despite retaining the lead. Figure 4A displays the observed variation; Appendix E reports each round, task-level agreement and changes in question outcomes, together with the recorded CLI and account differences.

4.4 Judge robustness

Judge choice preserves agent rankings and failure patterns. We re-score 442 interpretive answers from the first round with simplified descriptions using five judge configurations spanning Claude and GPT models, with the same reference facts and scoring prompt. Changing the judge produces no reversal in the agent ranking, with one tie between Luna and Terra. Claim interpretation remains the dominant failure category for five of the six agents under every judge. The benchmark judge agrees with the panel's median verdict on 89.8% of answers. Absolute success rates vary across judges, but agent rankings and dominant failure patterns remain consistent.

Repeated judgments are consistent. On a balanced 180-answer retest, the five configurations reproduce their own verdicts in 92.8% to 97.8% of cases. This checks repeatability on fixed answers, separately from the variation between agent runs. Appendix N reports judge configurations, per-judge results and the comparison between judge and agent model families.

4.5 Qualitative study

Expert review of selected tasks, questions and saved analyses from Papers 23 and 24 examines whether agents support their answers with an appropriate analysis. Opus 5 provides the strongest analysis in these reviewed cases, but still misses weak signals and necessary calculations. Figure 6 focuses on a thin bismuth crystal (Chen et al., 2024). The task links changes in electrical resistance under magnetic field and gate voltage to the charge carriers in the material. Oscillations in resistance reveal contributions from different carrier populations.

Weaker signals remain unresolved. Opus recovers the dominant electron-like contribution and useful carrier-density estimates, yet does not consistently resolve the two weaker hole-like contributions required by the reference. Most other agents stop at broad resistance trends or make incorrect conversions from oscillation frequency to carrier density. Correct answers about the number of carrier groups can therefore coexist with incomplete signal extraction. Missing these weaker signals leaves the analysis without a reliable account of the carrier populations behind the measured response.

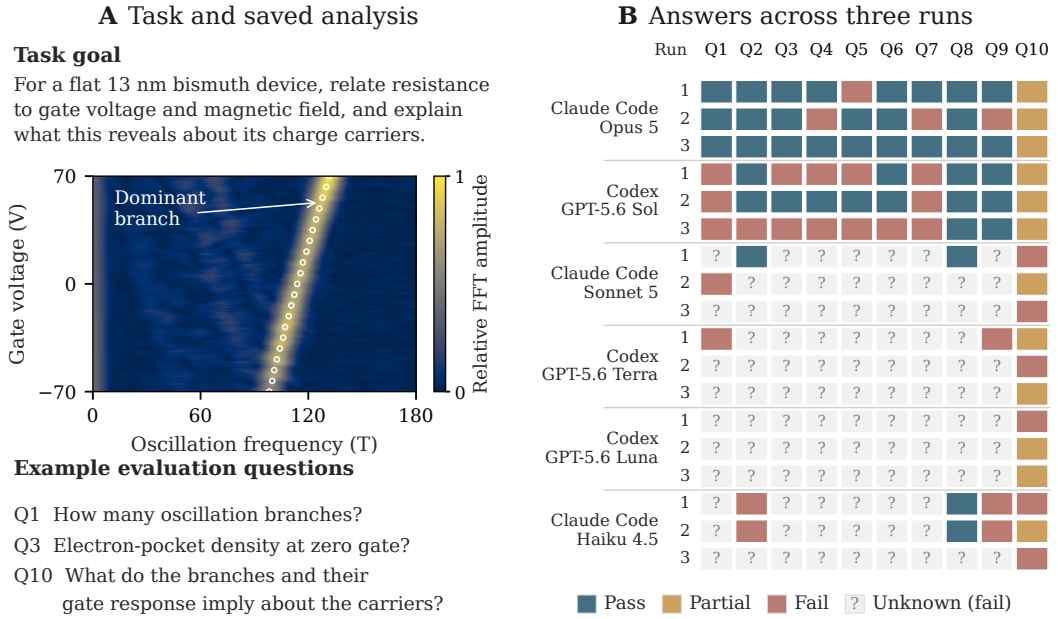


Figure 6: **Qualitative study of Paper 24, Task 04.** Opus tracks the dominant oscillation but leaves weaker carrier contributions unresolved. **A** Opus 5's third-round spectrum from its saved analysis; circles mark the frequencies it tracked. The task goal and example questions are abbreviated. **B** Question verdicts across agents and runs. Each row is a run and each column a question; Q10 tests the final interpretation. Appendix K gives processing details.

Necessary calculations are omitted. Across these cases, agents sometimes recognize a physical effect without performing the calculation needed to establish it. In the $HfTe_5$ task (Liu et al., 2025), some agents omit the conversion from measured resistivities to the difference in electron and hole densities. In a separate bismuth task, all six agents omit the required energy-gap fit in the reviewed analyses with simplified descriptions. More explicit task guidance prompts fitting attempts, but the resulting estimates remain incorrect in the reviewed runs.

Plausible explanations can lack quantitative support. Scientific interpretation requires a physical explanation that fits the measurement conditions. In the $HfTe_5$ case, agents sometimes invoke general caveats or acquisition details that obscure the central conclusion about carrier balance. In the bismuth task (Q10), Opus and Sol identify coexisting electron-like and hole-like surface carriers but leave out the quantitative comparison with earlier measurements needed to support that assignment. The required pocket sizes are not stated in the main text of earlier papers on bismuth. They must be calculated from values extracted from those papers' measurement figures and converted to a common basis. This need for figure analysis goes beyond general knowledge or text retrieval and helps explain why no agent fully answers Q10.

5 Conclusion

LEAP-Bench provides 64 experimental analysis tasks grounded in claims from 12 materials-science papers. Its research briefs, experimental inputs and reference questions connect task success to the quantities, reasoning and interpretations needed for a scientific conclusion. Across three evaluation rounds, correct intermediate answers can coexist with incomplete claim interpretations, and stable total scores can mask inconsistent task completion. Twelve tasks remain unsolved across all models and runs. The tasks and saved analyses provide a basis for future training environments. Feedback could identify missing evidence and guide agents to revise the analysis supporting a claim. Task-level evaluation would test whether training improves completion and consistency across runs.

AI USE STATEMENT

We used AI tools under human supervision and review to support the labor-intensive preparation of the benchmark. The authors provided prompts and guided revisions as agents helped interpret source papers, organize data, and prepare initial drafts of tasks and reference questions. AI tools also assisted with implementation checks, result verification and manuscript editing. A separate Claude Fable 5 judge evaluates interpretive answers, and five judge configurations assess scoring consistency and repeatability. Section 3.2 describes domain-expert review. The authors retain responsibility for the scientific decisions, final content and released artifacts.

REPRODUCIBILITY STATEMENT

We will release the construction pipeline, evaluation harness and analysis scripts, together with task packages, question specifications, reference answers, data provenance, scoring criteria and tolerances. The release will include agent and judge prompts, model and environment configurations, saved agent outputs and scoring records. These materials support rerunning the evaluation and recomputing the reported metrics from saved responses and verdicts. Appendix M describes construction and input provenance; Appendix L details the execution protocol and scoring. Appendix E records the repeated runs and their execution conditions, while Appendix N documents the judge-robustness analysis.

ETHICS STATEMENT

A shared benchmark for experimental analysis can help research groups compare agents on tasks that otherwise require substantial domain-specific setup. Its preserved inputs, analyses and reference questions also support inspection of how an agent arrived at an answer. A principal risk is that a fluent or correct answer may be used to justify a scientific conclusion despite errors in the underlying analysis. The saved artifacts enable such cases to be examined, and further expert evaluation can establish when an agent's analysis is suitable for scientific use. The benchmark uses experimental datasets released by the source papers' authors.

REFERENCES

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961, 2025. doi: 10.1038/s43588-025-00836-3. URL https://www.nature.com/articles/s43588-025-00836-3.

Anthropic. Claude Science, an AI workbench for scientists, is now available. https://www.anthropic.com/news/claude-science-ai-workbench, 2026. Accessed September 2026.

Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36–D42, 2013. doi: 10.1093/nar/gks1195.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous benchmarking of AI

agents with a scientific research suite. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b2ce9568dbb559aefc8c98ca5b5314ce-Abstract-Conference.html.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Laisi Chen, Amy X. Wu, Naol Tulu, Joshua Wang, Adrian Juanson, Kenji Watanabe, Takashi Taniguchi, Michael T. Pettes, Marshall A. Campbell, Mingjie Xu, Chaitanya A. Gadre, Yinong Zhou, Hangman Chen, Penghui Cao, Luis A. Jauregui, Ruqian Wu, Xiaoqing Pan, and Javier D. Sanchez-Yamagishi. Exceptional electronic transport and quantum oscillations in thin bismuth crystals grown inside van der Waals materials. *Nature Materials*, 23(6):741–746, 2024. doi: 10.1038/s41563-024-01894-0.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13936–13971. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.815. URL https://aclanthology.org/2024.findings-emnlp.815/.

Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7:23, 2015. doi: 10.1186/s13321-015-0068-4.

Ziyang Huang, Yi Cao, Ali K. Shargh, Jing Luo, Ruidong Mei, Mohd Zaki, Zhan Liu, Wyatt Bunstine, William Jurayj, Somdatta Goswami, Tyrel McQueen, Michael Shields, Jaafar El-Awady, Paulette Clancy, Benjamin Van Durme, Nicholas Andrews, William Walden, and Daniel Khashabi. Can coding agents reproduce findings in computational materials science? In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.00803.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10116. Curran Associates, Inc., 2024. doi: 10.52202/079017-0324. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13836f251823945316ae067350a5c366-Abstract-Datasets_and_Benchmarks_Track.html.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Jinyu Liu, Varsha Subramanyan, Robert Welser, Timothy McSorley, Triet Ho, David Graf, Michael T. Pettes, Avadh Saxena, Laurel E. Winter, Shi-Zeng Lin, and Luis A. Jauregui. Possible spin-triplet excitonic insulator in the ultraquantum limit of $HfTe_5$. *Physical Review Letters*, 135(4):046601, 2025. doi: 10.1103/bj2n-4k2w.

Wanhao Liu, Jiaqing Xie, Qian Tan, Weida Wang, Jue Wang, Ran Sun, Zhuo Yang, Wanli Ouyang, Lei Bai, Tianfan Fu, Lu Chen, Xin Chen, and Yuqiang Li. OmniMatBench: A human-calibrated multimodal reasoning benchmark across 19 materials science subfields, 2026. URL https://arxiv.org/abs/2605.29833. arXiv preprint arXiv:2605.29833.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Dániel L. Barabási, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zwezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodriques, Michaela M. Hinks, and Andrew D. White. Kosmos: An AI scientist for autonomous discovery, 2025. URL https://arxiv.org/abs/2511.02824. arXiv preprint arXiv:2511.02824.

Utkarsh Pratiush, Hiroshi Funakubo, Rama Vasudevan, Sergei V. Kalinin, and Yongtao Liu. Scientific exploration with expert knowledge (SEEK) in autonomous scanning probe microscopy with active learning. *Digital Discovery*, 4(1):252–263, 2025. doi: 10.1039/D4DD00277F.

Shi Qiu, Junyi Deng, Yiwei Deng, Haoran Dong, Jieyu Fu, Mao Li, Zeyu Li, Zhaolong Zhang, Huiwen Zheng, Leidong Bao, Anqi Lv, Zihan Mo, Yadi Niu, Yiyang Peng, Yu Tian, Yili Wang, Ziyu Wang, Zi-Yu Wang, Jiashen Wei, Liuheng Wu, Aoran Xue, Leyi Yang, Guanglu Yuan, Xiarui Zhan, Jingjun Zhang, Zifan Zheng, Pengfei Liu, Linrui Zhen, Kaiyang Li, Qichang Li, Ziheng Zhou, Guo-En Nian, Yunwei Xiao, Qing-Hong Cao, Linjie Dai, Xu Feng, Peng Gao, Ying Gu, Chang Liu, Jia Liu, Ming-xing Luo, Yan-Qing Ma, Liang-You Peng, Huichao Song, Shufeng Wang, Chenxu Wang, Tao Wang, Yi-Nan Wang, Chengyin Wu, Pengwei Zhao, and Hua Xing Zhu. PRBench: End-to-end paper reproduction in physics research, 2026. URL https://arxiv.org/abs/2603.27646. arXiv preprint arXiv:2603.27646.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=BsMMc4MEGS.

Tao Song, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Guoqing Zhang, Fei Zhang, Weiwei Shang, Yao Fu, Jun Jiang, and Yi Luo. A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 147(15):12534–12545, 2025. doi: 10.1021/jacs.4c17738.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56843–56873. PMLR, 2025. URL https://proceedings.mlr.press/v267/starace25a.html.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Märta A. Tschudin, David A. Broadway, Patrick Siegwolf, Carolin Schrader, Evan J. Telford, Boris Gross, Jordan Cox, Adrien E. E. Dubois, Daniel G. Chica, Ricardo Rama-Eiroa, Elton J. G. Santos, Martino Poggio, Michael E. Ziebel, Cory R. Dean, Xavier Roy, and Patrick Maletinsky. Imaging nanomagnetism and magnetic phase transitions in atomically thin CrSBr. *Nature Communications*, 15(1):6005, 2024. doi: 10.1038/s41467-024-49717-9.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02905.

Yonghao Weng, Liqiang Gao, Linwu Zhu, and Jian Huang. MatQnA: A benchmark dataset for multi-modal large language models in materials characterization and analysis, 2025. URL https://arxiv.org/abs/2509.11335. arXiv preprint arXiv:2509.11335.

Sifan Wu, Huan Zhang, Yizhan Li, Farshid Effaty, Hongyuan Mei, Amirreza Ataei, and Bang Liu. Seeing beyond words: MatVQA for challenging visual-scientific reasoning in materials science. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23599–23614. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1181. URL https://aclanthology.org/2026.findings-acl.1181/.

Yuetong Wu and Maojun Sun. AutoXRD: Autonomous LLM agents and comprehensive evaluation for powder diffraction analysis, 2026. URL https://arxiv.org/abs/2609.00070. arXiv preprint arXiv:2609.00070.

Zujie Xie, Zixuan Chen, Jiheng Liang, Xiangyang Yu, and Ziru Yu. LUMIR: an LLM-driven unified agent framework for multi-task infrared spectroscopy reasoning. *Analytica Chimica Acta*, 1382: 344857, 2026.

Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, George Michalopoulos, Yue Zhang, Ziyang Zhang, Mian Zhang, Zhiyu Chen, and Xinya Du. LMR-BENCH: Evaluating LLM agents' ability on reproducing language modeling research. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6164–6186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.314. URL https://aclanthology.org/2025.emnlp-main.314/.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024. doi: 10.1039/d3dd00188a.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. MatSciBench: Benchmarking the reasoning ability of large language models in materials science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 12868–12879, 2026. doi: 10.1145/3770855.3818888. URL https://doi.org/10.1145/3770855.3818888.

Zhen Zhang, Zhichu Ren, Chia-Wei Hsu, Weibin Chen, Zhang-Wei Hong, Chi-Feng Lee, Aubrey Penn, Hongbin Xu, Daniel J. Zheng, Shuhan Miao, Yimeng Huang, Yifan Gao, Weiyin Chen, Hugh Smith, Yaoshen Niu, Yunsheng Tian, Ying-Rui Lu, Yu-Cheng Shao, Sipei Li, Hsiao-Tsu Wang, Iwnetim I. Abate, Pulkit Agrawal, Yang Shao-Horn, and Ju Li. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature*, 647(8089):390–396, 2025. doi: 10.1038/s41586-025-09640-5.

Zhiwen Zhu, Shaoxuan Yuan, Quan Yang, Hao Jiang, Fengru Zheng, Jiayi Lu, and Qiang Sun. Autonomous scanning tunneling microscopy imaging via deep learning. *Journal of the American Chemical Society*, 146(42):29199–29206, 2024. doi: 10.1021/jacs.4c11674.

A BENCHMARK COMPOSITION

Table 2: **Benchmark composition.** The 12 papers cover different material systems and experimental measurements, with 64 tasks and 475 questions in total. Task specifications identify the available input files, including instrument measurements and any selected processed or supplementary inputs. Articles, evaluator specifications and reference answers remain separate from agent packages.

ID	System and phenomenon	Experimental inputs given to the agent	Tasks	Questions
P00	Amorphous-selenium ovonic threshold-switching selector: charge-triggered switching	Pulsed and DC electrical transport (I–V sweeps, oscilloscope switching transients); optical-absorption and photoemission spectra	9	62
P01	Kagome metal $CsTi_3Bi_5$: superconductivity and nematic order	Scanning tunneling microscopy topographs and $\mathrm{d}I/\mathrm{d}V$ maps (quasiparticle interference)	2	15
P05	Interfacial shear modulus and friction of 2D materials	Friction-force microscopy maps and force–distance curves; Raman spectra	3	25
P06	Atomically thin CrSBr: nanomagnetism and magnetic phase transitions	Scanning nitrogen-vacancy magnetometry image sequences; AFM topography	3	17
P11	Co_2MnGa: topological Weyl fermion lines and drumhead surface states	Angle-resolved photoemission (ARPES) volumes; magnetization and magnetotransport curves	6	37
P13	WS_2/MoS_2 heterostructure: interlayer hole and electron transfer	Two-dimensional electronic spectroscopy maps and time traces; absorption spectra	4	36
P17	Triangular-lattice magnet $NaRuO_2$: quantum disordered ground state	Neutron diffraction and inelastic neutron scattering; magnetization, heat capacity and transport curves	10	72
P20	Nano-columnar La_2NiO_4 thin-film cathodes: electrode performance	SEM, TEM and AFM images; electrochemical impedance spectra	4	30
P23	$HfTe_5$ in the ultraquantum limit: possible spin-triplet excitonic insulator	Pulsed-magnet magnetotransport records (longitudinal and Hall waveforms)	3	22
P24	Thin bismuth crystals grown inside van der Waals materials: transport and quantum oscillations	Magnetotransport and gate-sweep databases, high-field sweeps; AFM topography	6	51
P44	Boron-doped silicon nanowires in a micro-harvester: thermoelectric properties	Temperature-dependent I–V curves, Seebeck (open-circuit voltage vs. ΔT) measurements; SEM images	5	38
P56	MgB_2Se_4 spinels (B = Sc, Y, Er, Tm) as Mg-ion solid electrolytes	Powder X-ray diffraction, solid-state NMR spectra, impedance spectroscopy and electrochemical cycling	9	70
Total			64	475

Table 3: **Question counts by purpose and answer format.** Key-quantity questions are predominantly numerical; most claim-interpretation questions require interpretive answers. Rows give the probe target and columns the recorded answer format. Every reasoning-step and claim-interpretation question must pass under Equation 1.

Probe	Numerical	Counting	Categorical	Binary	Interpretive	Total
Key quantity	188	12	4	4	–	208
Reasoning step	2	7	88	91	6	194
Claim interpretation	–	–	5	–	68	73
Total	190	19	97	95	74	475

The inventory spans several forms of experimental analysis. Value extraction and model fitting account for 37 of the 64 tasks, while measurement comparison and pattern identification each contribute 11. Five tasks cover four less frequent analysis types. Table 2 lists the material systems

Table 4: **Task success and exclusive failure categories across three rounds.** Values are mean ± sample SD in percentage points, each with 64 tasks as the denominator. Unsolved tasks are assigned first to reasoning failure, then to interpretation failure, and otherwise to quantity failure. This priority defines an exclusive scoring partition, not a temporal account of the analysis. All runs use simplified descriptions; partial verdicts are non-passing under the strict rule. The four columns sum to 100% before rounding in each round.

Model	Strict success (%)	Reasoning failure (%)	Interpretation failure (%)	Quantity failure (%)
Opus 5	59.4 ± 7.2	9.4 ± 1.6	28.1 ± 4.1	3.1 ± 2.7
GPT-5.6 Sol	38.0 ± 0.9	13.5 ± 3.3	45.3 ± 2.7	3.1 ± 0.0
Sonnet 5	20.8 ± 3.3	30.7 ± 0.9	45.3 ± 3.1	3.1 ± 1.6
GPT-5.6 Terra	19.3 ± 3.9	31.2 ± 4.1	49.0 ± 7.2	0.5 ± 0.9
GPT-5.6 Luna	16.7 ± 3.3	34.9 ± 2.4	45.8 ± 2.4	2.6 ± 0.9
Haiku 4.5	0.5 ± 0.9	78.6 ± 2.4	19.8 ± 2.4	1.0 ± 0.9

Table 5: **Task success by analysis type across three rounds.** Values are mean ± sample SD of strict task-success percentages under simplified descriptions, with each column's task count as denominator. *Across measurements* combines evidence across modalities or measurements. *Other analyses* covers physical attribution, anomaly diagnosis, raw-data processing and contextual synthesis.

Model	Quantity extraction ($n = 20$)	Model fitting ($n = 17$)	Across measurements ($n = 11$)	Feature identification ($n = 11$)	Other analyses ($n = 5$)
Opus 5	75.0 ± 0.0	43.1 ± 9.0	45.5 ± 9.1	60.6 ± 22.9	80.0 ± 0.0
GPT-5.6 Sol	46.7 ± 2.9	31.4 ± 3.4	30.3 ± 5.2	27.3 ± 0.0	66.7 ± 11.5
Sonnet 5	36.7 ± 2.9	13.7 ± 6.8	6.1 ± 5.2	6.1 ± 5.2	46.7 ± 30.6
GPT-5.6 Terra	33.3 ± 10.4	13.7 ± 3.4	9.1 ± 0.0	15.2 ± 5.2	13.3 ± 11.5
GPT-5.6 Luna	20.0 ± 5.0	17.6 ± 10.2	12.1 ± 5.2	9.1 ± 0.0	26.7 ± 23.1
Haiku 4.5	1.7 ± 2.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

and measurements, and Table 3 relates the scientific purpose of each question to its answer format. Task type is visible to the agent. Reference chains contain 5 to 12 steps and define three complexity tiers, with six, 41 and 17 tasks; the chains and tiers are evaluator-only metadata.

B Full Task-Level Results

Tables 4 and 5 report three-round mean results and sample standard deviations. We assign each unsolved task to its first failed criterion in the order reasoning steps, interpretation and quantities. A task can fail several criteria; this priority rule assigns it to one category. The measurement breakdown separates failures of the quantity or reasoning criteria from interpretation failures when both criteria pass. The joint diagnostic in Table 6 applies the stricter condition that every intermediate answer passes. Appendix E reports each round, including success with partial verdicts accepted.

B.1 Interpretation after correct quantities and reasoning

In Section 4.5, expert review identifies gaps between agents' numerical results and their scientific interpretations. We examine this distinction quantitatively by asking how often an agent still fails a conclusion check when all of its quantity and reasoning answers pass.

Conditioning and aggregation. For each model and round, we select tasks for which every quantity and reasoning answer receives a pass, then count those with at least one non-passing interpretation answer. This condition requires zero quantity errors; the primary task-success rule permits one. The conditional percentage uses the selected tasks as its denominator within each round. Table 6 reports the mean and sample SD of these round-level percentages, alongside mean task counts.

Table 6: **Conclusion failures after fully correct quantity and reasoning answers.** The first three numeric columns are mean task counts over all three rounds. The second numeric column adds the all-reasoning-correct requirement to the first; the last two columns describe tasks in that subset with any non-passing conclusion answer. The rate is mean $\pm$ sample SD of the round-level conditional percentages, not a pooled ratio. Tasks with no quantity questions satisfy the quantity condition. †Haiku has no eligible task in round 3; its rate uses the two defined rounds, while its count means use all three. Partial verdicts count as non-passing.

Model	All quantity answers correct	+ All reasoning answers correct	Conclusion failure count	Conclusion failure (% of subset)
Opus 5	41.0	40.0	10.0	24.8 ± 10.7
GPT-5.6 Sol	38.7	35.3	17.3	49.2 ± 5.1
Sonnet 5	28.0	24.0	12.3	51.3 ± 9.9
GPT-5.6 Terra	32.3	27.0	17.7	65.1 ± 6.6
GPT-5.6 Luna	25.3	20.0	12.0	60.1 ± 9.0
Haiku 4.5	3.7	2.0	1.7	$83.3 \pm 23.6^{\dagger}$

Correct intermediate answers do not ensure complete conclusions. Passing every quantity and reasoning-step question does not ensure a complete account of what the measurements mean (Table 6). Among tasks with all these intermediate answers correct, Sol still misses at least one conclusion check on about half of tasks, and Terra on about two thirds, averaged across rounds. Opus also misses these checks on about one quarter of eligible tasks. Because these questions assess physical meaning, such as the mechanism behind an observed trend, intermediate-answer accuracy alone misses a substantive part of scientific task completion. Figure 5 provides a complementary breakdown of quantity, reasoning and interpretation failures by measurement type.

C Shared Failures and Analysis Types

The main results identify twelve tasks that remain unsolved across all models and rounds. Table 7 separates this persistent set from failures shared within an individual round. Figure 7 shows the complete question-score distributions by analysis type, complementing the strict task-success summary in Figure 4B. These views distinguish partial progress from completing the scientific task.

Analysis categories. Value extraction derives physical quantities from measurements; model fitting estimates parameters or scaling laws; measurement comparison combines evidence across techniques, samples or conditions; pattern identification locates or classifies features in signals and images.

Table 7: **Shared task coverage across configurations.** None, one and at least two give mean ± sample SD of task counts across three rounds, pairing the six independent configurations by repeat index. These three columns partition each row in each round. Never solved counts tasks unsolved in all 18 configuration–round combinations; it is not a round mean. All entries use strict success with simplified descriptions. Other analyses contains five tasks across four less frequent types.

Analysis type	Tasks	None per round	One per round	At least two per round	Never solved
Quantitative extraction	20	3.7 ± 0.6	3.7 ± 1.2	12.7 ± 0.6	2
Model fitting	17	6.7 ± 1.5	4.7 ± 0.6	5.7 ± 1.2	4
Across measurements	11	4.7 ± 0.6	3.3 ± 0.6	3.0 ± 1.0	4
Feature identification	11	3.7 ± 1.5	3.7 ± 1.5	3.7 ± 1.5	1
Other analyses	5	1.0 ± 0.0	0.3 ± 0.6	3.7 ± 0.6	1
All tasks	64	19.7 ± 2.3	15.7 ± 3.5	28.7 ± 4.0	12

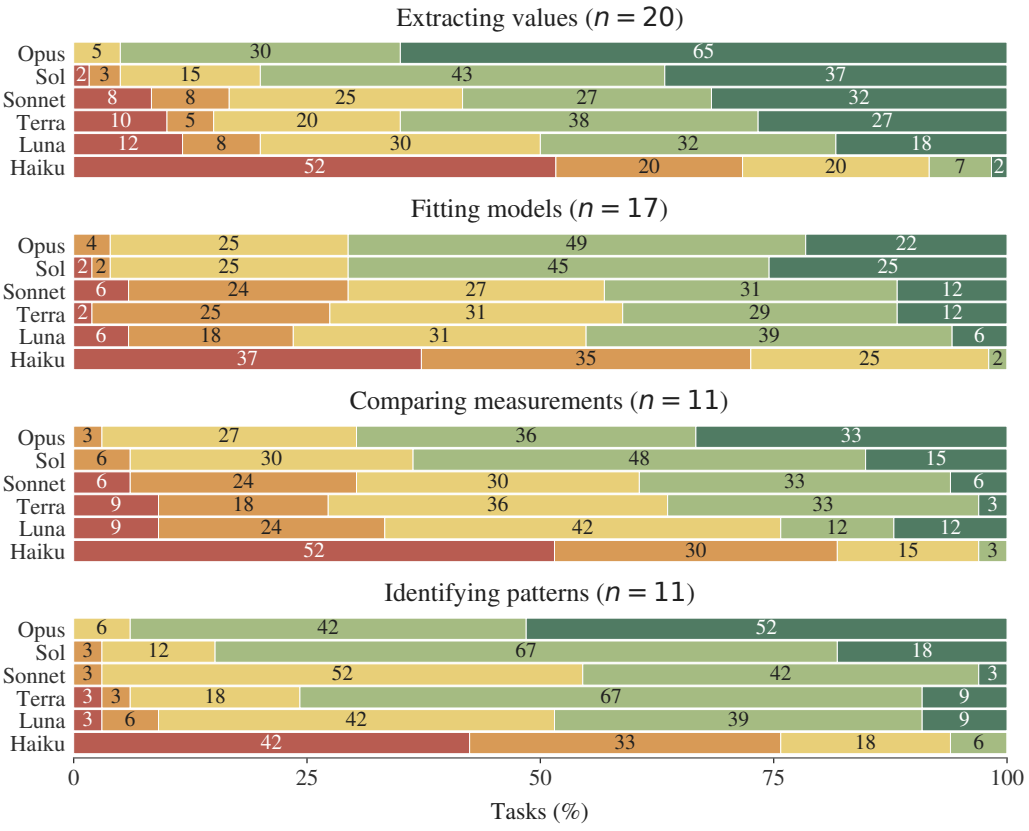


Figure 7: **Task-score distributions by required analysis.** Tasks with every question correct are less common in model fitting and measurement comparison than in value extraction. From red to dark green, segments give mean task percentages with $[0, 25\%)$, $[25, 50\%)$, $[50, 75\%)$, $[75, 100\%)$ and exactly 100% of questions correct across three rounds. Only pass verdicts count as correct; bars sum to 100% before rounding. These categories cover 59 tasks; Table 5 includes all 64.

D Detailed Results by Measurement Type

Figure 5 locates failures in the analysis and interpretation checks. The distributions below show how much of each task's required evidence agents recover, including progress that does not meet the task-success rule. Figure 8 groups the measurements into four families; Figures 9 and 10 show the score distributions and task outcomes for all six measurement types.

Measurement families and scores. Electrical transport is grouped with bulk magnetic and thermal measurements; microscopy and scanning probes are grouped with scattering and diffraction. Spectral measurements and electrochemistry form separate families. Each task is counted once within a family, and tasks requiring several measurement techniques can appear in multiple families. A task's score is the fraction of questions receiving a pass. The distributions average task percentages over the same three rounds with simplified descriptions.

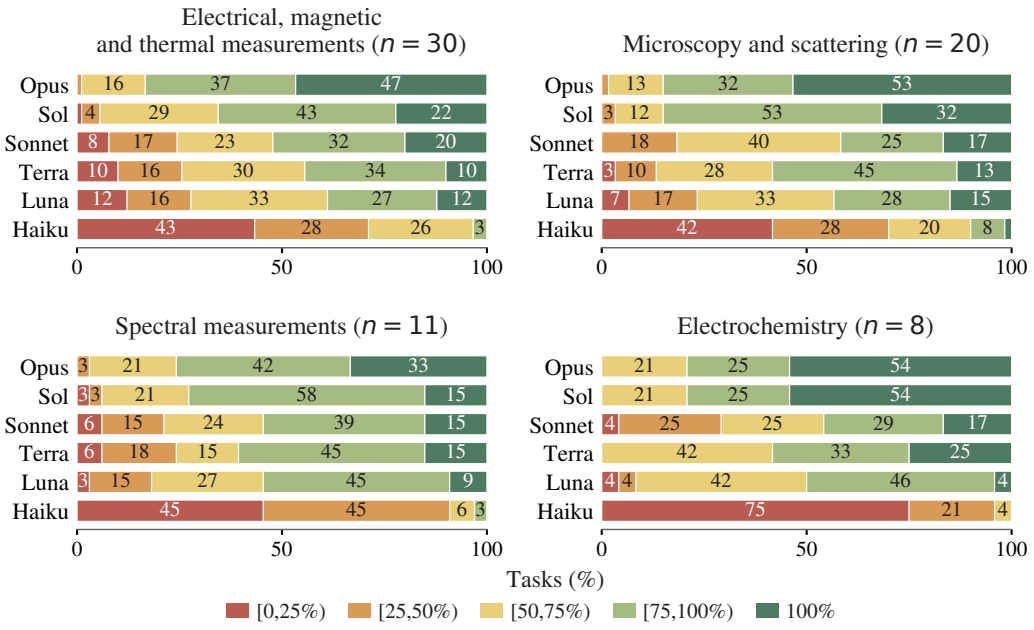


Figure 8: **Task-score distributions across measurement families.** Opus answers every question correctly least often on spectral tasks. Bars show mean task percentages across three rounds and sum to 100% before rounding. A task's score is the fraction of questions receiving a pass. Exactly 100% requires every question to pass; primary task success allows one quantity error. Tasks are deduplicated within each family and can appear in several families.

Spectral measurements challenge even the leading agent. Opus answers every question correctly on fewer spectral tasks than in the other three measurement families (Figure 8). These tasks analyze signals across energy, wavelength or resonance frequency. Figure 5 distinguishes failures in recovering that evidence from failures in interpreting it. Differences across groups also reflect their scientific objectives and task difficulty; the comparison does not isolate a causal effect of measurement type.

Failure counts and score distributions. In Figure 5, each of the n tasks in a measurement group is evaluated three times, giving $3n$ attempts. Both panels use all $3n$ attempts as their denominator. For example, Sol's $37/72$ in panel B for electrical measurements counts 37 attempts that pass the quantity and reasoning criteria but fail interpretation, out of 72 attempts. The quantity criterion permits one error, as in Equation 1. Table 6 uses the stricter subset in which every quantity and reasoning answer passes. The distributions below retain the full question scores and the exclusive task-outcome categories for each measurement type.

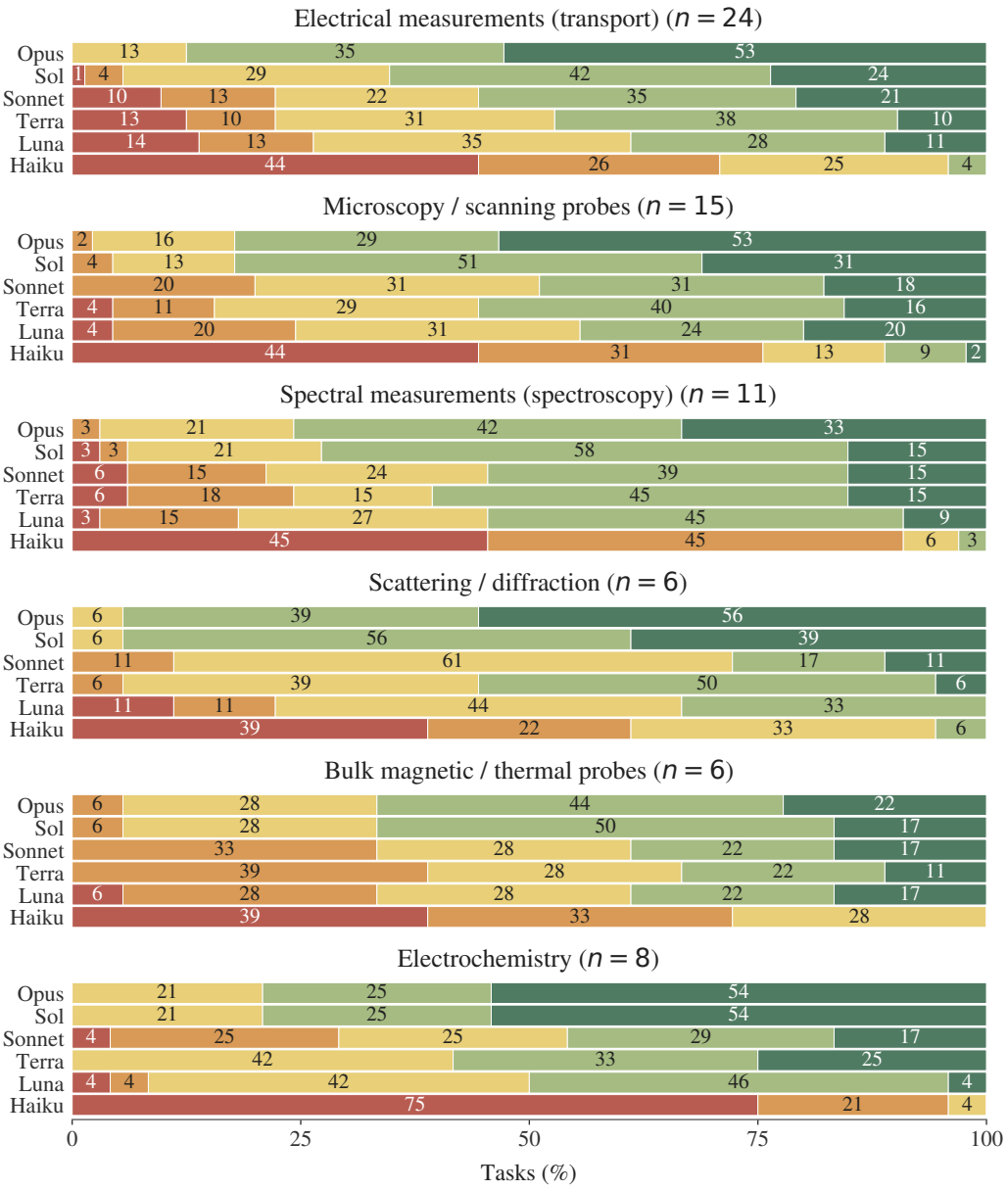


Figure 9: **Task-score distributions by measurement type.** Most spectral tasks contain an incorrect answer even for Opus. Scores are the fraction of questions receiving a pass. From red to dark green, score intervals are $[0, 25\%)$, $[25, 50\%)$, $[50, 75\%)$, $[75, 100\%)$ and exactly 100%. Bars show mean task percentages across three rounds and sum to 100% before rounding. Tasks using several techniques appear in each applicable group.

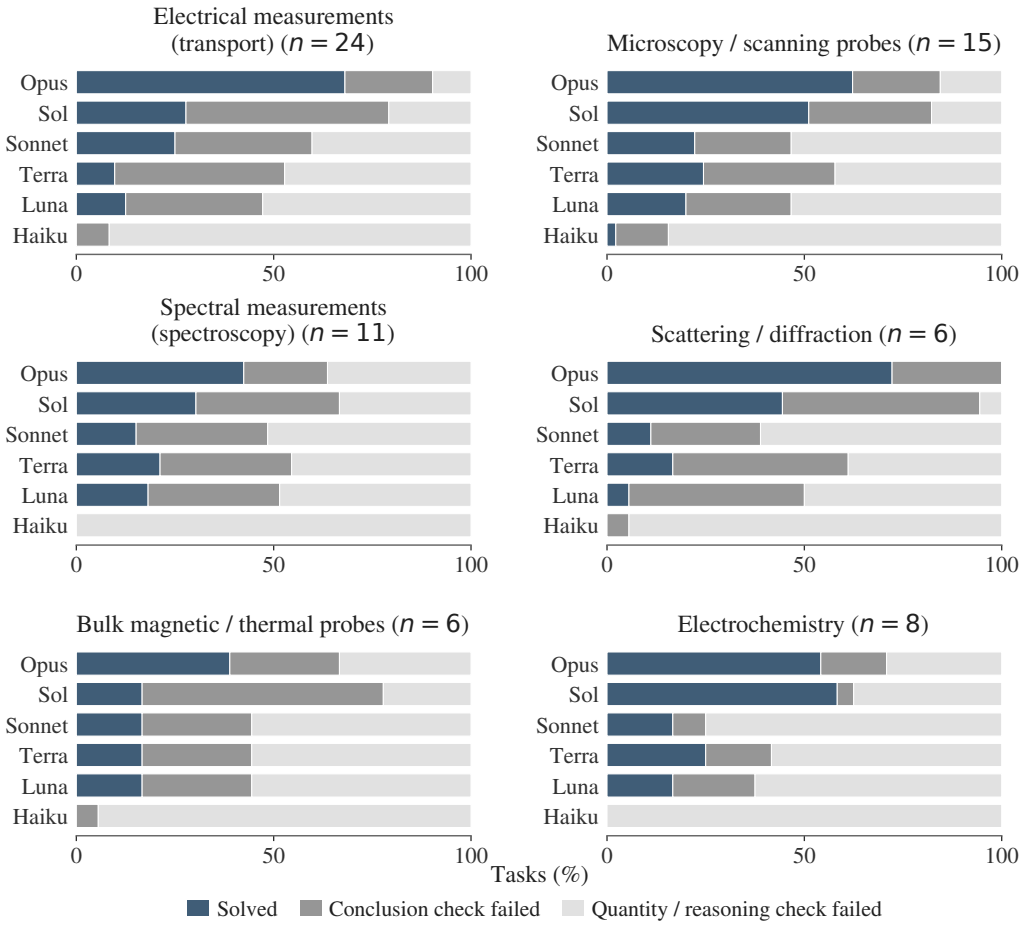


Figure 10: **Task success and failed checks by measurement type.** Sol often passes quantity and reasoning checks without completing the interpretation on electrical tasks. Bars show mean task percentages across three rounds and sum to 100%. Conclusion failures pass the quantity and reasoning criteria; the remaining failures miss at least one of these criteria. The rule permits one quantity error (Equation 1). Tasks using several techniques appear in each applicable group.

Table 8: **Task success across repeated runs.** Each cell counts tasks solved strictly in exactly zero, one, two or three of the three rounds. These four columns partition the 64 tasks. At least once is their union over repeats, not a single-run success rate. All rounds use simplified descriptions.

Model	0 rounds	1 round	2 rounds	3 rounds	At least once
Opus 5	16	11	8	29	48
GPT-5.6 Sol	28	12	11	13	36
Sonnet 5	40	11	10	3	24
GPT-5.6 Terra	43	10	6	5	21
GPT-5.6 Luna	48	6	4	6	16
Haiku 4.5	63	1	0	0	1

Table 9: **Task-success percentages in every repeat.** Each round evaluates the same 64 tasks with simplified descriptions. The upper panel uses strict success. The lower panel gives full credit to pass and partial verdicts for all question types before applying the same task rule. Means and sample SDs use the three round-level percentages, without pooling task attempts.

Strict task success				
Model	r1	r2	r3	Mean $\pm$ SD
Opus 5	65.6	60.9	51.6	59.4 ± 7.2
GPT-5.6 Sol	37.5	39.1	37.5	38.0 ± 0.9
Sonnet 5	21.9	23.4	17.2	20.8 ± 3.3
GPT-5.6 Terra	15.6	23.4	18.8	19.3 ± 3.9
GPT-5.6 Luna	20.3	14.1	15.6	16.7 ± 3.3
Haiku 4.5	1.6	0.0	0.0	0.5 ± 0.9

Task success with all partials accepted				
Model	r1	r2	r3	Mean $\pm$ SD
Opus 5	76.6	75.0	71.9	74.5 ± 2.4
GPT-5.6 Sol	76.6	79.7	75.0	77.1 ± 2.4
Sonnet 5	42.2	31.2	45.3	39.6 ± 7.4
GPT-5.6 Terra	56.2	53.1	50.0	53.1 ± 3.1
GPT-5.6 Luna	57.8	43.8	46.9	49.5 ± 7.4
Haiku 4.5	4.7	6.2	1.6	4.2 ± 2.4

E Repeated Evaluations and Run Stability

The repeated evaluation uses three complete rounds on the same 64 tasks and 475 questions for each configuration. Each round contains one session per source paper and configuration, yielding 216 paper-level runs and 8,550 question verdicts in total. Table 9 gives each round's task success and the three-round mean and sample standard deviation. Table 10 gives the corresponding ordered failure partition. All rates are computed within a round before averaging; the standard deviation of three-round success uses denominator $3 - 1$. The three rounds are repeated evaluations of the same benchmark, so their task outcomes are not independent samples of new scientific tasks.

Execution conditions. Task and question files and the two initial phase instructions match across rounds, and all recorded model and package checks pass. Codex uses CLI version 0.154.0 and `xhigh` effort throughout; its first round uses API-key authentication, while the second and third use ChatGPT-account authentication. Claude Code uses versions 2.1.278 and 2.1.280 in the first round and 2.1.280 in the later rounds. The execution record also documents a different Claude account for the third round. Nine first-round Claude sessions contain an additional recorded call for recovery or completion, whereas every second- and third-round session completes with the two standard phase calls. These differences are retained when interpreting the observed variation; the evaluation does not isolate an account effect or variation due only to model sampling.

Table 10: **Success and exclusive failure percentages in every repeat.** Every row uses 64 tasks and strict success with simplified descriptions. Unsolved tasks are assigned to reasoning, interpretation or quantity failures, in that priority order. These are scoring categories and do not describe execution order. The four categories partition each round; percentages may not sum to 100% after rounding.

Model	Round	Strict success	Reasoning failure	Interpretation failure	Quantity failure
Opus 5	r1	65.6	9.4	23.4	1.6
	r2	60.9	7.8	29.7	1.6
	r3	51.6	10.9	31.2	6.2
GPT-5.6 Sol	r1	37.5	12.5	46.9	3.1
	r2	39.1	10.9	46.9	3.1
	r3	37.5	17.2	42.2	3.1
Sonnet 5	r1	21.9	31.2	45.3	1.6
	r2	23.4	29.7	42.2	4.7
	r3	17.2	31.2	48.4	3.1
GPT-5.6 Terra	r1	15.6	29.7	53.1	1.6
	r2	23.4	35.9	40.6	0.0
	r3	18.8	28.1	53.1	0.0
GPT-5.6 Luna	r1	20.3	32.8	43.8	3.1
	r2	14.1	34.4	48.4	3.1
	r3	15.6	37.5	45.3	1.6
Haiku 4.5	r1	1.6	76.6	20.3	1.6
	r2	0.0	78.1	21.9	0.0
	r3	0.0	81.2	17.2	1.6

Repeated task completion. Table 8 counts how often each task succeeds across three rounds. A task solved at least once enters the union of solved tasks; a task solved in all three enters their intersection. These descriptive counts do not assume an oracle that selects a successful answer or an automatic retry policy. We separately compare each pair of rounds to measure changes in task status and in the pass status of quantity, reasoning and interpretation answers (Table 11).

Interpretation changes can alter task completion. The correctness of the final scientific interpretation varies across repeated runs. For Sol, the pass status of claim-interpretation questions changes on 19% to 26% of questions between pairs of rounds, compared with 3% to 6% for reasoning-step questions. Opus's lower third-round success likewise coincides with fewer passing claim-interpretation answers, while its quantity and reasoning-step accuracies change little. These are changes in scored outcomes of repeated analyses, not evidence that a particular execution setting caused the difference. The recorded account, authentication and CLI differences are summarized in Appendix E.

F Effect of Task Guidance

Some agents solve more tasks with simplified descriptions. In the first-round comparison, the original descriptions provide more guidance on the analysis, yet Sonnet 5 and Opus 5 solve more tasks with the simplified descriptions (Table 12). The other four configurations solve more tasks with the original descriptions. Each condition uses one saved run per paper, with different Claude Code versions, so the comparison does not isolate the effect of task wording.

Similar success rates do not mean the same tasks are solved. The first-round comparison also shows that similar total success can involve different tasks under each description. Opus 5 solves three more tasks overall with simplified descriptions, but six tasks are solved only with the original descriptions and nine only with the simplified descriptions. Sol likewise has different successful tasks in the two conditions despite a small difference in total success. The two description formats thus reveal different failures within the same task set. A repeated comparison with matched execution settings could test which differences persist across runs.

Table 11: **Agreement and changes between pairs of repeats.** Task agreement is the percentage of the 64 tasks with the same strict solved/unsolved status; Tasks changed gives the complementary count. The last three columns are question-level pass/non-pass flip percentages within quantity ($n = 208$), reasoning ($n = 194$) and conclusion ($n = 73$) probes. Partial and fail are both non-pass; a partial-to-fail change is not counted as a flip. Agreement can be high because both repeats fail the same task. All runs use simplified descriptions.

Model	Pair	Task agreement	Tasks changed	Quantity flips	Reasoning flips	Conclusion flips
Opus 5	r1–r2	82.8	11	11.5	3.1	12.3
	r1–r3	79.7	13	10.1	4.1	16.4
	r2–r3	78.1	14	11.1	5.2	17.8
GPT-5.6 Sol	r1–r2	76.6	15	12.5	3.1	19.2
	r1–r3	81.2	12	12.5	5.2	20.5
	r2–r3	70.3	19	12.5	6.2	26.0
Sonnet 5	r1–r2	76.6	15	18.3	13.9	17.8
	r1–r3	82.8	11	23.6	15.5	17.8
	r2–r3	75.0	16	21.6	16.0	24.7
GPT-5.6 Terra	r1–r2	82.8	11	15.9	9.3	11.0
	r1–r3	81.2	12	14.4	7.7	13.7
	r2–r3	85.9	9	13.9	8.8	8.2
GPT-5.6 Luna	r1–r2	84.4	10	18.8	11.3	16.4
	r1–r3	89.1	7	16.3	11.3	16.4
	r2–r3	95.3	3	18.8	10.3	11.0
Haiku 4.5	r1–r2	98.4	1	22.6	33.5	8.2
	r1–r3	98.4	1	13.9	33.0	12.3
	r2–r3	100.0	0	21.2	33.5	12.3

Table 12: **Original versus simplified task descriptions.** Some agents solve more tasks with simplified descriptions. Original descriptions provide more analytical guidance; simplified descriptions leave more choices to the agent. Both versions use the same 64 tasks, data, questions and references. Columns give solved-task counts and distinguish tasks solved with both versions from those solved with only one. Each condition uses one saved run per paper; CLI versions are documented in Appendix L. These values use the first evaluation round.

	Solved tasks		Tasks solved with each version		
Configuration	Original	Simplified	Both	Original only	Simplified only
Haiku 4.5	3	1	0	3	1
Sonnet 5	12	14	7	5	7
Opus 5	39	42	33	6	9
GPT-5.6 Luna	15	13	6	9	7
GPT-5.6 Terra	18	10	6	12	4
GPT-5.6 Sol	27	24	16	11	8

G Results by Paper

Table 13 gives the number of tasks solved in each paper and round. The relative difficulty of a material system can persist even when particular tasks change status. For example, Opus 5 solves five tasks in P00 in every round, while its count on P56 decreases from seven to five. These counts locate variation across the deposited scientific analyses without treating papers as interchangeable tasks.

H Results by Reference-Analysis Complexity

Table 14 reports mean success across the three rounds by reference-chain length. The task and question counts show the different group sizes underlying these rates, complementing the analysis-

Table 13: **Per-paper task successes in each repeat.** Every model cell reports strict solved-task counts as r1/r2/r3, with the paper's task count in n. The last row sums each round over all 12 papers. Opus, Sonnet and Haiku denote Claude configurations; Sol, Terra and Luna denote GPT-5.6 configurations. All runs use simplified descriptions.

ID	System (primary measurement)	n	Opus	Sol	Sonnet	Terra	Luna	Haiku
P00	a-Se OTS selector (electrical transport)	9	5/5/5	1/2/2	1/2/2	0/0/0	0/0/0	0/0/0
P01	$CsTi_3Bi_5$ kagome metal (STM/QPI)	2	1/1/1	2/2/2	0/0/0	2/2/2	2/1/1	0/0/0
P05	2D interfacial friction (friction-force AFM)	3	2/3/2	1/1/1	2/1/2	0/2/1	1/1/1	0/0/0
P06	Few-layer CrSBr (scanning NV magnetometry)	3	3/2/2	2/2/2	1/1/1	0/1/1	0/1/1	0/0/0
P11	Co_2MnGa Weyl lines (ARPES)	6	6/5/5	3/3/3	1/2/1	3/2/3	3/1/2	0/0/0
P13	WS_2/MoS_2 charge transfer (2D spectroscopy)	4	1/1/0	1/0/2	2/0/1	0/1/2	1/1/1	0/0/0
P17	$NaRuO_2$ triangular magnet (neutron, thermodynamics)	10	5/4/4	3/3/2	1/2/0	2/1/0	1/0/0	0/0/0
P20	La_2NiO_4 nano-columnar cathodes (SEM, EIS)	4	2/2/0	1/1/1	0/0/0	0/0/0	0/0/0	0/0/0
P23	$HfTe_5$ ultraquantum limit (pulsed-field transport)	3	2/1/1	0/1/1	1/2/1	0/0/1	2/0/1	0/0/0
P24	Thin Bi in vdW crystals (magnetotransport)	6	4/4/4	2/3/1	0/0/0	0/0/0	0/0/0	1/0/0
P44	B-doped Si nanowires (thermoelectric transport)	5	4/4/4	2/2/2	3/3/1	1/1/0	2/1/1	0/0/0
P56	MgB_2Se_4 spinel electrolytes (XRD, NMR, EIS)	9	7/7/5	6/5/5	2/2/2	2/5/2	1/3/2	0/0/0
All papers		64	42/39/33	24/25/24	14/15/11	10/15/12	13/9/10	1/0/0

Table 14: **Mean task success by reference-analysis complexity.** Values are mean strict task-success percentages over three simplified rounds. Each row's task count is the denominator. Tiers count steps in the hidden reference chain; model abbreviations follow Table 13.

Complexity	Tasks	Questions	Opus	Sol	Sonnet	Terra	Luna	Haiku
Tier 1 ($\leq$6 steps)	6	37	66.7	38.9	27.8	11.1	22.2	0.0
Tier 2 (7–9 steps)	41	297	62.6	38.2	22.8	19.5	14.6	0.8
Tier 3 ($\geq$10 steps)	17	141	49.0	37.3	13.7	21.6	19.6	0.0

type breakdown in Figure 4B. Reference-chain length describes the expected analysis, so this comparison does not isolate its causal effect on task difficulty.

I Sensitivity to the Task-Success Rule

The primary rule evaluates whether a task's answers jointly support its scientific claim. Requiring complete reasoning and interpretation reflects that goal. The one-quantity allowance permits a limited error among the intermediate estimates, alongside the question-specific numerical tolerances. It is applied uniformly, including the seven tasks with at most one quantity question. We hold this rule fixed in the main comparisons and use variants to examine answer completeness.

Table 15 compares three-round mean success under the primary rule, all questions correct, at least 80% of quantities correct, and acceptance of partial verdicts. The partial-acceptance variant gives full question credit to both pass and partial verdicts across all question purposes. The primary comparison continues to require full answers. These diagnostics describe sensitivity to answer completeness, with each rule applied to the same saved verdicts.

Table 15: **Mean task success under alternative scoring rules.** Values are mean percentages of 64 tasks over three rounds with simplified descriptions. The first three rules require every reasoning and interpretation question to pass. The quantity threshold in the second column is 80%; Strict allows at most one non-passing quantity answer. All partials accepted uses the strict task rule after counting both pass and partial as correct for every probe target; it is not half credit. Per-round strict and partial-accepted results appear in Table 9.

Model	All questions correct	Quantities $\geq$80% correct	Strict	All partials accepted
Opus 5	46.9	47.9	59.4	74.5
GPT-5.6 Sol	28.1	29.2	38.0	77.1
Sonnet 5	18.2	18.2	20.8	39.6
GPT-5.6 Terra	14.6	14.6	19.3	53.1
GPT-5.6 Luna	12.5	14.6	16.7	49.5
Haiku 4.5	0.5	0.5	0.5	4.2

Table 16: **Task wording and data-access conditions.** Original runs have higher question accuracy for every configuration, while differences in task success vary. Values are percentages of 64 tasks or 475 questions, using the same scoring rules. *Original* and *Simplified* supply the experimental inputs. *No data* uses original wording without the files and requests best estimates; the other conditions permit unknown answers. These values use the first evaluation round.

	Task success			Question accuracy		
Model	Original	Simplified	No data	Original	Simplified	No data
Haiku 4.5	4.7	1.6	3.1	39.4	29.3	43.6
Sonnet 5	18.8	21.9	6.2	70.7	69.9	52.4
Opus 5	60.9	65.6	9.4	89.3	88.2	57.5
GPT-5.6 Luna	23.4	20.3	10.9	77.3	67.2	57.3
GPT-5.6 Terra	28.1	15.6	4.7	78.5	70.1	56.8
GPT-5.6 Sol	42.2	37.5	6.2	84.8	82.3	62.1

J Question-Level Diagnostics

First-round diagnostic cohort. The comparisons in this section retain the first simplified round alongside the unrepeated original-description and no-data runs. Their numbers describe that diagnostic cohort, while the main results and repeated-evaluation tables aggregate all three simplified rounds.

Task wording and data access. We compare the simplified condition with original task descriptions and with a no-data condition to examine the contribution of task wording and prior knowledge (Table 16). Original and simplified runs use the same data and questions; original descriptions give more analytical direction. Question accuracy is higher with original descriptions for every configuration; task success increases for some configurations and decreases for others.

No-data runs use original task descriptions, the same questions and reference answers, and best-estimate prompts. They receive neither measurement nor supplementary files. Simplified runs use task descriptions with less analytical guidance and permit unknown responses. Their comparison with no-data runs combines differences in file access, wording and response policy, which should be retained when interpreting the results.

Question answerability. Without data, configurations correctly answer many reasoning-step questions while completing few tasks. Table 17 reports no-data agreement across configurations and mean accuracy. Tables 18 and 19 give the full probe-target and answer-format breakdowns. Binary and categorical answers are more often correct without files than numerical and interpretive answers, separating prior answerability from the task's combined requirements.

A brief-coverage review flagged two questions for further examination. `PAPER_24_TASK_04_Q07` concerns the lowest Landau index at an extreme gate setting, and `PAPER_24_TASK_05_Q04`

Table 17: **Question-level results without measurement data.** Most reasoning-step questions are answered correctly by at least four no-data configurations, while numerical and interpretive answers show lower agreement. The first two percentage columns count questions answered correctly by at least four configurations or by none. Accuracy columns average six configurations, and *Difference* is simplified minus no-data accuracy in percentage points. Probe targets and answer types describe the same question set; conditions follow Appendix J.

Category	n	≥4/6 correct, no data (%)	0/6 correct, no data (%)	No data (%)	Simplified (%)	Difference (pp)
By what the question probes						
Key quantity	208	31	32	36.0	63.8	+27.8
Reasoning step	194	87	4	85.8	82.6	-3.2
Claim interpretation	73	21	48	26.9	40.0	+13.0
By answer type						
Numerical	190	26	34	31.8	64.3	+32.5
Counting	19	79	11	76.3	59.6	-16.7
Categorical	97	87	3	84.4	83.7	-0.7
Binary	95	94	1	91.9	84.7	-7.2
Interpretive	74	16	51	22.7	36.5	+13.7
All questions	475	52	23	54.9	67.8	+12.9

Table 18: **Question accuracy by probe target.** Opus 5 and Sol show the largest interpretation differences between the simplified and no-data conditions. Values are first-round pass rates within each question purpose; partial verdicts count as incorrect. Pairs compare the same configuration across the two conditions in Appendix J. Bold marks the highest simplified value in each group.

		Key quantity (n=208)		Reasoning step (n=194)		Claim interpretation (n=73)	
Harness	Model	data	no data	data	no data	data	no data
Claude Code	Haiku 4.5	12.5	21.2	50.0	75.8	21.9	21.9
	Sonnet 5	66.3	31.2	87.1	82.5	34.2	32.9
	Opus 5	**85.6**	36.1	**96.4**	91.2	**74.0**	28.8
Codex CLI	GPT-5.6 Luna	65.4	39.9	82.5	86.1	31.5	30.1
	GPT-5.6 Terra	72.1	41.3	84.0	87.1	27.4	20.5
	GPT-5.6 Sol	80.8	46.2	95.9	92.3	50.7	27.4

Table 19: **Question accuracy by answer format.** Numerical and interpretive questions separate the conditions more clearly than binary and categorical questions. Values are pass rates within each format, with partial verdicts counted as incorrect. Pairs compare simplified and no-data results for the same configuration (Appendix J); bold marks the highest simplified value in each group. These values use the first evaluation round.

		Numerical (n=190)		Counting (n=19)		Categorical (n=97)		Binary (n=95)		Interpretive (n=74)	
Harness	Model	data	no data	data	no data	data	no data	data	no data	data	no data
Claude Code	Haiku 4.5	11.1	17.4	26.3	57.9	52.6	71.1	52.6	85.3	16.2	17.6
	Sonnet 5	68.9	26.8	52.6	73.7	89.7	80.4	85.3	89.5	31.1	28.4
	Opus 5	**85.8**	31.6	**78.9**	84.2	**96.9**	94.8	97.9	91.6	**73.0**	24.3
Codex CLI	GPT-5.6 Luna	66.3	35.3	57.9	78.9	83.5	83.5	86.3	93.7	25.7	27.0
	GPT-5.6 Terra	72.6	37.4	68.4	78.9	83.5	85.6	87.4	94.7	24.3	14.9
	GPT-5.6 Sol	81.1	42.6	73.7	84.2	95.9	90.7	**98.9**	96.8	48.6	24.3

concerns the onset field of a splitting. Both are quantity questions retained in the scored set. Excluding these two questions leaves all observed strict task-success outcomes unchanged.

K A Saved Analysis and Its Claim Interpretation

Figure 3 source. The Method example uses the $HfTe_5$ measurements of Liu et al. (2025) in `PAPER_23_TASK_02`. The task body is reproduced verbatim from the simplified task, with

shortened context and selected input paths; the displayed questions correspond to Q02, Q03 and Q07. For the plot, we calibrate opposite-polarity TDMS recordings at 0.5 K and 10 K using the logged gains and sample geometry, interpolate descending-field traces to a common 0.1 T grid, and antisymmetrize the transverse resistivity. No smoothing is applied. The interpretation reference additionally uses longitudinal resistivity to obtain Hall conductivity by tensor inversion.

Qualitative spectrum. Figure 6 is reconstructed from the measurement array and processing code saved during Opus 5's third-round analysis phase. FFT amplitudes use one global normalization across gate voltages. The carrier-density question Q3 uses $n = eF/h$ without a spin multiplier.

A saved magnetic-image analysis. A separate task uses five deposited nitrogen-vacancy magnetometry files from Tschudin et al. (2024). Its seven evaluation questions concern the evolution of a magnetic transition in a CrSBr bilayer. The saved analysis below comes from the simplified GPT-5.6 Terra run on `PAPER_06_TASK_02`; Section 4.5 examines a bismuth transport task. The analysis, numerical results and judge record support the observations summarized below.

Analysis decisions. The agent averages the forward and backward `Norm` images, recording their agreement as a noise check. It selects a common threshold of $\mathrm{Norm} > 1.10$ to track high-signal morphology and explicitly treats this as a relative measure. The area above this threshold decreases across all five scans. The agent also measures shrinking connected regions and increasing fragmentation, and uses these observations in its subsequent interpretation.

Answers and reference checks. For the area-fraction questions, the agent normalizes its high-signal counts to the first scan. Its first and last fractions are 1.0 and 0.1266, compared with references of 0.88 and 0.0, each with an absolute tolerance of 0.15. Both pass. Its three step answers also pass, identifying a decreasing area, localized boundary motion and the initial fully magnetized configuration. These checks establish agreement for the queried outputs under the declared tolerances.

Interpretive differences. The mechanism answer describes boundary motion and excludes a coherent whole-flake switch, but omits the reference's nucleation step; the saved judge marks it partial. For pinning, the agent links irregular boundaries and residual bright regions to spatially heterogeneous pinning. The reference links smooth incremental motion to weak pinning, and the saved judge marks the answer as a contradiction. This example illustrates a difference between passing intermediate-answer checks and completing a paper-grounded interpretation. It does not independently validate the agent's full analysis or adjudicate the physical pinning mechanism.

L Experimental Setup Details

Agent configurations. Across the three simplified rounds, agents use Claude Code versions 2.1.278 and 2.1.280 with Haiku 4.5, Sonnet 5 or Opus 5, and Codex CLI version 0.154.0 with GPT-5.6 Luna, Terra or Sol. Claude Code uses the product's default reasoning effort; Codex uses `xhigh`. The comparison uses each product's configured effort setting. The runner disables user and project settings and memory, supplies the same task and phase prompts within each condition, and records model usage and trajectories. Appendix E records the round-specific CLI, authentication and recovery conditions. Claude permits subagents pinned to the evaluated model, while Codex disables them.

Task-guidance comparison. The original and simplified descriptions retain the same task identifiers, types, required files, questions and reference answers. Simplified task descriptions provide less guidance on the analysis in the title, context and task body. The comparison pairs saved outcomes on the same 64 tasks with one run per paper and condition. Codex uses version 0.154.0 and `xhigh` effort throughout. Original Claude Code runs use versions 2.1.261, 2.1.269 and 2.1.278; simplified runs use 2.1.278 and 2.1.280. These conditions limit causal attribution to the task wording alone.

Run isolation and phase instructions. The sandbox exposes the run's input files as read-only and provides writable output space. Reference answers and the rest of the project are outside the run. Network access is limited by a logged proxy to the endpoints required by the agent service. Before scoring, the evaluator checks the package fingerprint, records input integrity, verifies served-model records and audits the trajectory. Trajectory flags are retained for further inspection.

The fingerprint covers the task view, question view and phase prompts; data-file consistency is checked separately. Both phases use the same mounted run directory. Phase instructions require analysis

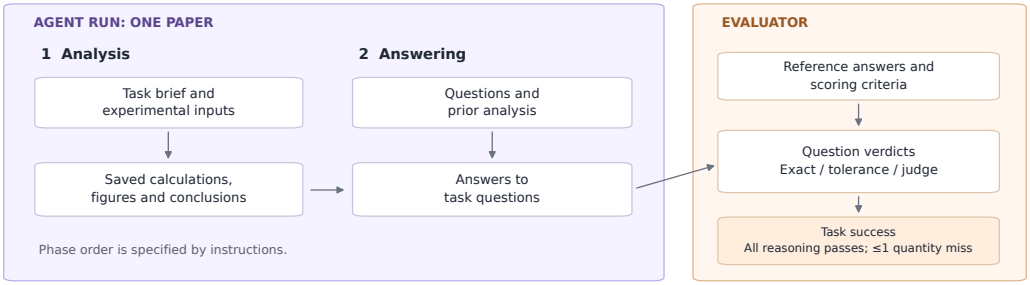


Figure 11: **Two-phase evaluation.** Agents analyze the inputs and save their results before answering evaluation questions in the same session. Deterministic comparisons and an independent language-model judge score the answers, which jointly determine task success.

Table 20: **Runtime, token use and estimated cost.** Mean API-equivalent cost ranges from $0.27 to $24.80 per paper in the simplified condition. *Call time* sums CLI-call durations across both phases; time and token values are medians over 12 papers. Input tokens include cache reads and Claude subagent usage. Cost columns give the mean per paper and total, excluding construction and judging. Appendix L describes the accounting. These values use the first evaluation round.

			Call time	Input tokens	Output tokens	Estimated cost (USD)	
Harness	Model	Effort	(min)	(M)	(k)	per paper	12 papers
Claude Code	Haiku 4.5	default	9.2	3.8	58	0.94	11
	Sonnet 5	default	30.3	24.0	248	8.50	102
	Opus 5	default	56.2	28.2	213	24.80	298
Codex CLI	GPT-5.6 Luna	xhigh	11.1	6.2	69	0.27	3
	GPT-5.6 Terra	xhigh	14.2	4.0	48	1.81	22
	GPT-5.6 Sol	xhigh	15.0	5.0	56	4.71	57

before opening questions and answers based on saved outputs without reopening measurements. Each run retains the integrity checks, serving-model records and trajectory.

Answer scoring. Each question stores its comparison rule, including exact matching and absolute, relative or logarithmic numerical tolerances. If x is the response and x^* the reference, the tolerance checks are $|x - x^*| \leq \tau_{\text{abs}}$, $|x - x^*| \leq |x^*|\tau_{\text{pct}}/100$, or $|\log_{10} x - \log_{10} x^*| \leq \tau_{\log}$, respectively; the logarithmic rule requires positive values. Categorical comparisons are case-insensitive where specified. Scoring also checks any accepted alternate references.

Interpretive answers use a separate Claude Fable 5 call in an empty working directory with tools disabled. Its input consists of the public question, expected answer, relevant evaluator scoring criteria, agent response and fixed rubric. It does not receive the source paper, measurement data, analysis code or output figures. The response records a fact ledger, verdict and rationale. Contradicting a reference fact is grounds for failure; omissions or qualification of the central conclusion can receive a partial verdict. The scorer uses the declared verdict, and the judge inputs, fact ledger and rationale are saved for review. Appendix N reports completed cross-judge agreement, repeatability and task-level robustness tests on the first evaluation round.

Usage accounting. For the first simplified round, Table 20 reports the sum of CLI-call durations across both phases, including repeated calls and excluding waits between them. Claude usage is reconstructed from unique response IDs in the saved main-session and subagent transcripts; Codex uses recorded incremental usage. Saved API prices determine estimated input, cache and output costs. These estimates exclude construction and judging and describe API-equivalent usage costs.

Table 21: **Construction statistics.** The retained pipeline produces 64 tasks and 475 verified questions from 108 inventoried claims. Gap resolutions are counted per task; *Files* counts distinct supplementary inputs. Review rounds and question outcomes refer to the current retained construction. *Partial briefs* retain additional wording needed to specify the intended analysis.

			Gaps closed				Suff.	Questions (blind verification)				Brief	
ID	Claims	Tasks	Folded	Derived	Rescoped	Files	rounds	Authored	Confirmed	Corrected	Dropped	Rounds	Partial
P00	12	9	3	3	1	2	1	62	57	5	0	2	0
P01	6	2	0	0	0	0	1	15	15	0	0	1	0
P05	12	3	2	1	6	1	1	26	23	2	1	1	0
P06	8	3	2	4	2	4	2	17	14	3	0	1	1
P11	8	6	2	1	7	1	1	37	37	0	0	1	0
P13	6	4	4	3	1	1	1	36	35	1	0	1	0
P17	10	10	4	3	1	2	1	72	71	1	0	1	2
P20	11	4	2	0	0	0	1	30	28	2	0	1	0
P23	8	3	0	3	1	1	1	22	21	1	0	1	0
P24	8	6	2	10	8	4	2	51	51	0	0	1	1
P44	8	5	5	2	2	1	1	38	38	0	0	1	0
P56	11	9	4	3	1	1	1	70	68	2	0	1	1
All	108	64	30	33	30	18	–	476	458	17	1	–	5

M Construction Pipeline Details

The four conceptual stages in Section 3.2 comprise the implementation steps in Table 22. Paper understanding includes paper analysis and data inventory; task formulation includes task generation and sufficiency review; evaluation design includes question authoring and answer verification; benchmark assembly includes projection, brief preparation, review and packaging. Scoring is a separate evaluation step. Each step has explicit input permissions and output schemas. Authoring and review use fresh sessions; the evaluator and public-text passes within task generation share a session. Simplified and no-data packaging preserve the original task descriptions and reference answers.

Projection and validation. The specification defines eleven check families: numerical reference leakage (V1), procedural hints (V2), assumed conclusions (V3), evaluator references (V4), allowed fields (V5), projection fidelity and ordering (V6), length limits (V7), answer-format completeness (V8), tolerance policy (V9), question clarity and relevance (V10), and supplementary-input contents and provenance (V11). Numerical leakage checks include values associated with other tasks in the same paper, with exemptions for genuine measurement settings. The mechanical verifier checks field sets, copied values and structural requirements. Semantic conditions, including whether wording implies a conclusion or supplies an analysis recipe, also require model review.

Tolerance policy. A reference records the basis of its numerical tolerance. For values stated in text or captions, the default uses 1.5 times the reported uncertainty or twice the half-unit of the last printed digit, with a 25% provenance cap. Values calculated from stated inputs use propagated uncertainty with a 30% cap; figure readings use axis resolution with a 50% cap. The policy permits a wider absolute band when justified by stated experimental uncertainty. Decade-scale quantities use a log-base-ten band, normally 0.3 and at most 0.5. Authoring and review apply these tolerance requirements; the scorer uses the stored comparison and retains any policy warnings. Of the 209 numerical and counting questions, 122 derive their references from text, 28 from captions, 41 from figure readings and 18 from calculations based on stated values. Four questions also accept documented alternatives to their primary reference values.

Question admission. Each question must state what it probes and identify the relevant sample, condition, region or comparison. When a quantity has multiple conventions, such as gap versus twice the gap or full versus half width, the question specifies the intended convention; a documented alternative can be accepted in the reference. Enumerations supply their options and numerical questions specify their units. References include a source anchor and should concern the task's primary evidence. At most two questions per task may concern supporting evidence, with a justification. Authoring and verification review scientific relevance, possible ambiguity and the consistency of the tolerance with the source. The stage contract calls for ambiguous or unsupported questions to be revised or dropped.

Table 22: **Construction steps and principal artifacts.** Authoring and review link published claims to tasks with specified inputs and reference answers. Each row records an implementation step's inputs, outputs and review role. File and schema checks enforce structural requirements, while model reviews assess scientific sufficiency, wording and reference consistency.

Step	Main inputs	Outputs	Review or separation
1 Paper analysis	Article, supplement and available evaluator references	Context; claim inventory, source anchors and conventions	Agent-facing and evaluator sections recorded separately
2 Data inventory	Deposited files and paper	File/sample inventory, claim coverage and missing metadata	Instrument files inspected separately from processed-data references
3 Task generation	Inventories and paper; evaluator references	Task specifications, gap resolutions and supplementary inputs	Evaluator specification precedes the public-text pass in one session
3b Sufficiency review	Tasks, deposited inputs and paper	Sufficiency decision, annotation checks and revision instructions	Fresh session; annotation re-derivation precedes reading provenance
4 Question authoring	Tasks, paper and evaluator references	Questions, reference answers, anchors and comparison rules	Question-purpose and tolerance requirements
4b Answer verification	Paper, public task fields and questions	Confirmed, corrected or dropped questions; regeneration flags	Fresh session; paper-based answer reconstruction
5 Projection	Task and question specifications	Agent and reviewer views	Field/value checks and specified semantic checks
5b Simplified brief	Public task and question text; data inventory	Revised task title, context and body	Inputs, task type and identifiers retained
5c Brief review	Brief and inventory, then original task and questions	Coverage judgments and revision instructions	Analysis plan written before reading the questions
6 Packaging	Public views and selected data	Run package and fingerprint	Content and data-consistency checks
7 Scoring	Answers and evaluator question/task records	Verdicts and judge logs	Separate tool-free judge for interpretive answers

Gap resolution. The preferred order is folded, derived, then rescoped. A folded requirement becomes part of the analysis the agent performs. A derived input supplies an annotation or constant, with its reconstruction recorded in a separate evaluator-side provenance file. A rescoped task limits the requested conclusion to what the available evidence supports. Tolerances should not be widened to compensate for missing information. If an expert annotation is available, construction can use it; otherwise it attempts a documented reconstruction. The 93 retained task-gap records correspond to 54 distinct paper/gap identifiers because several tasks can use the same missing information.

Input content and data availability. The source paper provides the target claims and reference answers. Agent packages use the experimental files selected in each task specification; article PDFs, evaluator references and author analysis scripts are retained separately. The selected deposits contain instrument measurements and some processed outputs. In P56, these include NMR panels labeled with peak positions, XRD refinement exports and impedance-fitting results. Other tasks use supplementary annotations reconstructed from figures or data renderings. Per-task file lists document the inputs used for evaluation.

Expert review and construction checks. Expert review of an earlier release containing five papers, 30 tasks and 236 questions produced twelve findings that informed the current construction requirements. The present inventory contains 108 claims, of which 64 become tasks. All retained tasks passed the final model-based sufficiency review, with at most two repair rounds. Answer verification confirmed 458 of 476 questions, corrected 17 and dropped one, yielding 475 questions with five

to ten per task. Current expert review examines selected tasks, questions and saved analyses from Papers 23 and 24, as reported in Section 4.5.

Model provenance. Construction is configured with Claude Fable 5, and serving-model usage is retained in the run records. Some task-generation envelopes also contain usage for other models. The separate interpretive-judge records identify Fable 5 as the serving model.

Construction counts. Table 21 records the retained inventories and reviews. The 93 task-level gap resolutions comprise 30 folded, 33 derived and 30 rescoped records across 48 tasks, using 18 supplementary files. There are 54 distinct paper/gap identifiers because several tasks can share a requirement. Two papers required a second sufficiency review, following a region-mask revision and a restriction to the measured field range, respectively. One paper required a brief revision, and five briefs retain extra wording needed to specify the intended analysis. The rounds and question outcomes describe the current retained construction.

N JUDGE ROBUSTNESS

Evaluation cohort. We test judge sensitivity on the first round with simplified descriptions. Of the 475 questions per agent, 401 use deterministic scoring and 74 require a language-model judge, including 68 claim-interpretation and six reasoning questions. Across six agents, 442 of the 444 responses require a judge; two unknown answers receive a failure directly under the existing scoring rule. We hold the deterministic verdicts, references and task-success rule fixed and replace only the verdicts supplied by the language-model judge.

Judge configurations and repeated scoring. The panel uses five configurations of four distinct models. Claude Fable 5, Fable 5.1 and Opus 5.5 use `max` reasoning effort, and GPT-6 Astra uses `max` and `xhigh`. The benchmark reference is Claude Fable 5 at default effort. All panel configurations receive the scorer's identical question, reference facts, criteria, agent answer and fact-ledger rubric, with tools and external information access disabled. Serving-model records are checked against the requested configurations. Using seed 20260925, we sample 60 answers from each official verdict category. Each configuration judges the resulting 180 answers twice. The balanced retest measures consistency when each judge evaluates an unchanged answer again.

Table 23: **Judge agreement and repeatability.** Agreement with the benchmark's default-effort Fable 5 judge uses all 442 re-scored answers. Repeat agreement uses the same 180 answers evaluated twice by each configuration, balanced across the three official verdict categories. Both agreement columns require exactly matching pass, partial or fail verdicts; κ_w is quadratic-weighted Cohen's kappa against the benchmark judge.

Judge	Effort	Agreement with benchmark judge (%)	κ_w	Repeat agreement (%)
Claude Fable 5	`max`	90.5	0.90	92.8
Claude Fable 5.1	`max`	86.9	0.87	93.3
Claude Opus 5.5	`max`	85.1	0.83	97.8
GPT-6 Astra	`max`	77.6	0.71	97.2
GPT-6 Astra	`xhigh`	77.4	0.70	97.8

Agreement and task-level conclusions. Table 23 reports exact three-way agreement and quadratic-weighted Cohen's κ_w, coding fail, partial and pass as 0, 1 and 2. Across the five configurations, Fleiss' κ is 0.69 and ordinal Krippendorff's α is 0.75. The official verdict matches the median of the five panel verdicts on 89.8% of responses ($\kappa_w = 0.88$). Table 24 shows how judge choice affects task scores. Opus 5 and Sol remain first and second throughout; the only change to the ordering is a tie between Luna and Terra under Fable 5 at `max` (Kendall's τ_b against the official scores: 0.966 to 1.000). Interpretation remains the largest assigned failure category for every agent except Haiku, whose failures concentrate in reasoning.

Table 24: **Task outcomes under alternative judges.** Each column re-scores the same 64 tasks per agent in the first simplified round. The upper block gives strict task success (%); the lower block gives counts of unsolved tasks assigned to reasoning / interpretation / quantity failures, in that priority order. Each unsolved task appears in one category according to this scoring priority. The categories do not indicate when an analysis fails. Only interpretive verdicts change; all deterministic verdicts and the task rule remain fixed. The default column is the benchmark judge used in the main results.

Agent	Fable 5 default	Fable 5 max	Fable 5.1 max	Opus 5.5 max	Astra max	Astra xhigh
Task success (%)						
Opus 5	65.6	70.3	73.4	64.1	37.5	37.5
GPT-5.6 Sol	37.5	43.8	46.9	39.1	20.3	18.8
Sonnet 5	21.9	23.4	26.6	28.1	15.6	14.1
GPT-5.6 Luna	20.3	20.3	25.0	25.0	12.5	10.9
GPT-5.6 Terra	15.6	20.3	23.4	20.3	7.8	7.8
Haiku 4.5	1.6	3.1	4.7	4.7	0.0	0.0
Failure counts: reasoning / interpretation / quantity						
Opus 5	6/15/1	5/12/2	5/10/2	5/15/3	7/32/1	7/32/1
GPT-5.6 Sol	8/30/2	7/27/2	7/24/3	7/29/3	8/43/0	8/44/0
Sonnet 5	20/29/1	19/29/1	19/27/1	19/25/2	20/34/0	20/35/0
GPT-5.6 Luna	21/28/2	21/27/3	21/25/2	20/27/1	21/34/1	21/35/1
GPT-5.6 Terra	19/34/1	17/32/2	17/30/2	18/31/2	19/39/1	19/39/1
Haiku 4.5	49/13/1	49/12/1	49/11/1	49/11/1	49/15/0	49/15/0

Differences in answer completeness. The panel configurations assign fail verdicts to 21.0% to 24.0% of answers, while pass rates range from 22.4% to 43.4%. Thus, similar aggregate fail rates coexist with different thresholds between pass and partial, which affect strict task success. For Opus 5, task success ranges from 37.5% to 73.4% across the panel; its leading rank is preserved throughout. The agent ordering and dominant failure categories remain consistent across the tested judges.

Judge family and agent family. For each response, we compare the mean ordinal verdict from the three Claude configurations with that from the two GPT configurations. The difference in this contrast between Claude-agent and GPT-agent responses is -0.013 on the 0–2 scale, with an item-bootstrap 95% interval of $[-0.107, 0.084]$ from 2,000 resamples. This analysis finds no detectable preference for the judge's own model family on that measure. The study evaluates model sensitivity under one fixed rubric and one evaluation round; the two GPT configurations share a model.

O LIMITATIONS

The current evaluation covers 12 selected papers and their deposited data, with variation in input processing and task composition. Each configuration is evaluated in three rounds; recorded CLI and account differences limit attributing the observed variation to model sampling alone. The original-description and no-data conditions have one run per paper. The no-data condition also changes task wording and response instructions. Section 4.4 tests the stability of interpretive scoring across alternative judges. Priorities for extending the evaluation include broader expert review, controlled repeats, matched data-access controls, and human and specialist-agent baselines.