How Is Automated Research Evaluated? A Survey of Benchmarks and Evaluation Practices

Liulei Zhang* **Dejing Zhou*** **Chuyue Huang** **Guanhua Chen**
Yutong Yao **Lidia S. Chao** **Chi Man Vong** **Derek F. Wong†**
NLP^{2}CT Lab, Department of Artificial Intelligence
University of Macau, Macau, China
nlp2ct.{liulei,chuyue,guanhua,yutong}@gmail.com
{mc55091,lidiasc,cmvong,derekfw}@um.edu.mo

Abstract

Automated research systems support literature synthesis, ideation, experiments, writing, and peer review, but their evaluation is dispersed across tasks, benchmarks, and studies that are difficult to compare directly. We review this literature from the perspective of evaluation design and evidence, covering six targets: literature synthesis, research ideation, executable workflows, scholarly writing and communication, automatic peer review, and end-to-end research. We compare task construction, evidence sources, evaluators, and scoring procedures to explain the capabilities assessed by different designs. Our synthesis highlights three recurring lessons: output checks, process checks, and human studies provide complementary information; evaluator calibration is specific to the property being assessed; and resource budgets and attempt selection are integral to interpreting performance comparisons. We identify diagnostic evaluation designs and documented gaps in supporting evidence, and translate these comparisons into reporting and audit recommendations for specific evaluation settings. The survey helps readers navigate existing evaluations, select appropriate benchmarks, and design subsequent studies.

1 Introduction

Research agents search papers, formulate proposals, run experiments, write, and review manuscripts. Correct computations, useful literature comparisons, and supported criticisms require different observations. Interpreting their benchmark scores therefore requires understanding the task, available evidence, and scoring procedure.

Existing surveys and perspectives organize automated research around the scientific method, workflow stages, autonomy, applications, and system architectures (Zhang et al., 2025a; Luo et al., 2025; Chen et al., 2025c; Zheng et al., 2025; Wei et al., 2025; Xie et al., 2025; Ren et al., 2026). Our review brings together the evaluation literature within this landscape: benchmarks, datasets, execution environments, human studies, metrics, and audit protocols. We compare their experimental designs as well as reported scores.

Across six evaluation targets, we distinguish questions requiring different evidence, such as a review's usefulness to its author versus its detection of a known error. Figure 1 maps representative works by target and claims supported by their core evaluation designs.

The survey makes three contributions. First, it *organizes and reviews evaluation studies*, explaining their tasks and assessment procedures. Second, it *compares evaluation designs and synthesizes findings*, connecting evidence, scoring, and calibration to assessed capabilities and support gaps. Third, it develops *setting-specific recommendations* for benchmark selection, result interpretation, and evaluation design. The appendix records protocols and reporting details.

2 Scope and Review Method

We review evaluation of automated research workflows, their artifacts, and human interaction through explicit tasks or inspectable scoring procedures. Standalone retrieval, single-paper question answering, generic coding, and autonomous wet-lab systems are outside the main coverage; selected adjacent tasks contribute distinctive designs to core comparisons.

Search and selection. We searched Google Scholar, arXiv, Semantic Scholar, and the ACL Anthology iteratively alongside writing, with back-

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2913–2951
November 6-10, 2026 ©2026 Association for Computational Linguistics

ward and forward citation tracing. The cutoff was May 10, 2026; descriptions of some studies incorporate subsequent revisions.

Queries combined target and evaluation terms (Appendix B). This selective review places representative comparisons in the main text and additional designs and details in the appendix. Cross-target studies are classified by their primary evaluation question.

Analysis fields. We compare the *target* (capability, artifact, or stage), *evidence source* (judgment materials and observations), *evaluator* (person, model, or procedure), *metric* (scoring and aggregation), and *validity boundary* (applicability conditions and identified claim–evidence gaps). Evidence-source analysis considers both provenance and support for output judgments. Related RAG studies examine evidence-based output calibration (Chen et al., 2025a) and multi-granularity retrieval for verification (Chen et al., 2026). We use these fields to check evidence and scoring against authors' stated claims, distinguish observations from proxies, and examine calibration, identifying conclusions supported under the reported conditions.

Five claim-strength levels organize interpretation. *L1: Surface plausibility* concerns artifact style or appearance. *L2: Task correctness under specified evidence* concerns success under defined inputs and checks. *L3: Process-grounded competence* requires criterion-checked process evidence explaining execution or decisions, beyond stored logs. *L4: Human- or utility-calibrated performance* relates results to defined populations' judgments or outcomes, distinguishing preference, factual assessment, and practical benefit. *L5: Scientific contribution* concerns advances supported by domain-appropriate novelty, methods, results, interpretation, and validation.

Cross-stage interpretation separates local results, verified overall completion, and stage interactions; process records explain interactions but are not universally required for reporting observable final outcomes.

Classifying representative works. Figure 1 classifies core evaluation designs and main experiments; supplementary process analyses, human ratings, and scorer audits do not independently determine placement. Works without a single-level summary of equally central designs are omitted. Table 1 and Appendix B detail selected classifications and protocols.

Most selected works assess task correctness (L2). BLADE, AstaBench (E2E), and MLR-Bench check analysis decisions or research steps (L3); RE-Bench and Scideator center expert-performance or user-judgment comparisons (L4).

3 Target-Specific Evaluation

We compare representative task designs and synthesize findings and open questions, interpreting specific protocols and reported results.

3.1 Scientific Literature Synthesis Evaluation

Literature synthesis turns multiple papers into an organized account of a research area or an answer to a research question. Evaluation must therefore specify who selects the sources, what the system produces, and how that output is checked. Long surveys, comparison tables, taxonomies, and question-driven answers expose different parts of this task; they are complementary evaluation settings, not successive stages of a single benchmark.

Survey generation and content assessment. AutoSurvey retrieves papers, builds an outline, and writes and revises individual sections (Wang et al., 2024b). Its evaluation separates citation support from coverage, structure, and relevance. Importantly, its citation recall measures support among sentences already containing citations, rather than recovery of a human bibliography. SurGE instead uses 205 human surveys and a shared retrieval pool to compare reference coverage, citation accuracy, structure, and content (Su et al., 2026). Its expert ranking study calibrates structure and content scores; it does not validate the separate citation judgments, which use titles and abstracts. Thus, similarly named citation metrics can have different denominators and evidence access. SurveyGen makes another important condition explicit: its 120-survey experiment distinguishes topic-only generation, automatic retrieval, and writing with human-provided references and outlines (Bao et al., 2025). The last condition tests writing with supplied organization, not autonomous source selection.

Structured synthesis. Comparison tables require choosing useful dimensions and filling them with paper-specific evidence. ArxivDIGESTables evaluates these operations separately: schema recall measures recovery of human comparison columns, while cell evaluation supplies the human schema

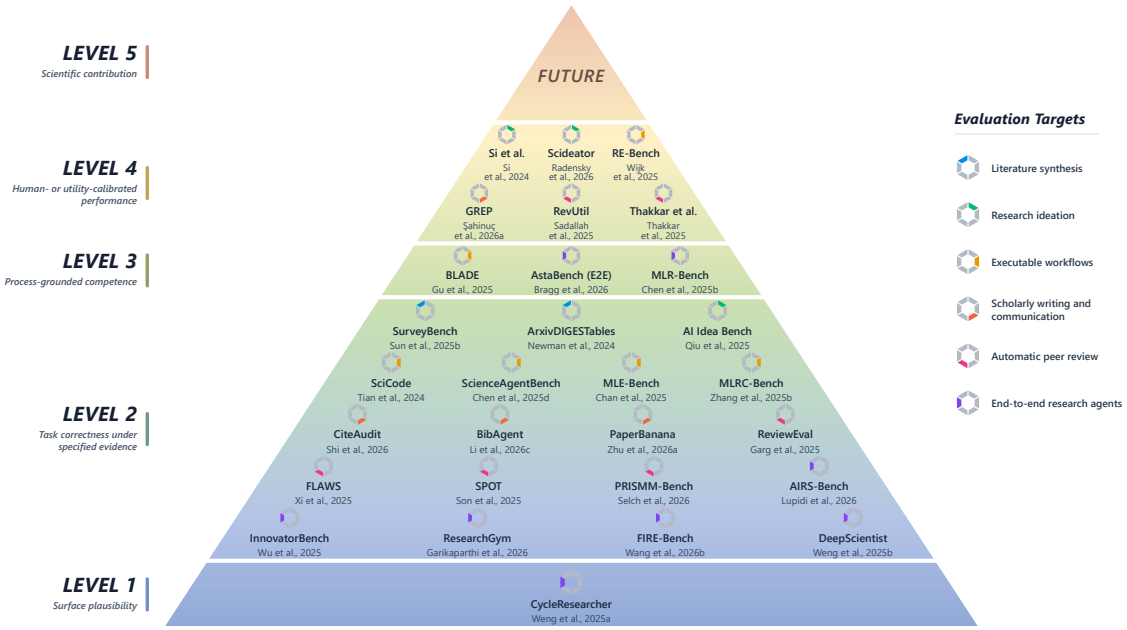


Figure 1: Representative works by claim strength. Highlighted hexagon segments indicate primary evaluation targets. AstaBench (E2E) denotes only the suite's end-to-end component. Level 5 indicates a direction for further evaluation development.

to isolate information extraction (Newman et al., 2024). Human checks cover 50 tables for column alignment and 30 for cell matching. Additional context mainly improves partial cell matches; generated columns absent from the reference are not significantly less useful in the reported study. Reference recovery and usefulness are therefore separately observable. TaxoBench similarly distinguishes finding expert-cited papers from organizing a supplied paper set into a hierarchy (Zhang et al., 2026b). Its reported structural controls show that semantic path scores respond strongly to tree depth and category wording, so raw alignment differences require interpretation alongside these controls.

Question-driven synthesis and information use. OpenScholar evaluates citation-bearing scientific answers using expert questions, rubrics, and preferences; its expert comparison covers 108 questions and 16 evaluators (Asai et al., 2024). ResearchQA scales a related task by mining questions and scoring requirements from surveys across 75 fields, with expert validation in eight fields (Yifei et al., 2025). These designs differ in how information needs are obtained and how widely human calibration extends. SurveyBench instead asks whether an already generated survey supplies the information needed for general and topic-specific quizzes (Sun et al., 2025b). Across 20 topics, an LLM answers from retrieved survey passages, and unsupported answers receive zero. DeepResearch scores 4.42/5 on content but 1.97/10 on topic-specific questions; LLM×MapReduce-V2 scores 4.20 and 3.19, respectively. This reversal shows why content ratings and supported answerability should both be inspected.

DeepSurvey-Bench adds a rubric for information, scholarly communication, and research guidance (Zhang et al., 2026a). It reports human calibration on 40 generated surveys and a separate evidence audit sampled from high-scoring surveys. That audit describes evidence quality within the selected group. More generally, Hwang et al. (2026) find that system-level agreement with experts can conceal weaker instance-level agreement in scientific long-answer evaluation.

Synthesis. The most informative designs separate source selection, organization, and use. Shared retrieval pools and supplied-outline conditions help identify which part of the task a comparison tests. Separate schema and cell evaluations similarly make an unsuccessful table easier to diagnose. ArxivDIGESTables' useful non-reference columns motivate assessing alternative structures alongside reference alignment; SurveyBench's differing content

and answerability rankings show why covered information must also be usable for answering questions.

Evaluation should match intended use: coverage and organization for field overviews, supported answerability for questions, and citation checks with explicit evidence access and denominators for source-grounded claims. Calibration must follow these distinctions. Expert agreement on content, structure, or preference does not validate separate citation or individual-answer scoring. Reporting these dimensions separately would reveal which component accounts for a system's strengths or failures.

3.2 Research Ideation and Proposal Evaluation

Research ideas are evaluated before their proposed experiments have produced results. Novelty asks how an idea differs from existing work; feasibility concerns whether its steps can be carried out; value concerns the expected importance of the contribution. Existing protocols approach these questions through expert judgments, comparison with published ideas, retrieved prior work, and evidence-bearing proposal reports.

Expert judgments and published references. Si et al. (2024) compare human ideas, automatically selected AI ideas, and human-reranked AI ideas in seven prompting-based NLP topics. Their study involves 104 distinct researchers and 49 proposals per condition, with blinded ratings after style normalization. At the idea level, AI proposals receive higher novelty scores than human proposals, while feasibility differences are not significant. This is a comparison of the specified generation and selection processes: AI uses extensive sampling and retrieval, whereas each human contributes one idea. AI Idea Bench asks a different question (Qiu et al., 2025). It supplies inspiration papers and a topic for 3,495 target papers, then evaluates generated motivations and experimental plans through target-paper matching and pairwise comparisons. Topic alignment, recovery of a published idea, and expert-perceived novelty consequently have different reference objects, even when all are used to evaluate ideation.

Novelty relative to prior work. Literature-Grounded Novelty Assessment makes the comparison set explicit (Shahid et al., 2025). It retrieves candidate papers, reranks them by purpose, mechanism, evaluation, and application, and judges novelty against the top ten using expert examples. On 32 held-out ideas, its reported accuracy is 0.81 and agreement with expert labels is $\kappa = 0.59$. A separate 58-case study tests only known non-novel ideas and should not be read as overall classification accuracy. Navigating Ideation Space instead separates problem, method, and findings representations, providing component-level links to prior work (Shen et al., 2026). Its novelty score correlates with 93 expert-rated ideas, but the reported Spearman correlation of 0.384 is close to GPT-5 Mini's 0.385. Scideator brings related facet-based reasoning into an interactive tool (Radensky et al., 2026). Its user study finds higher creativity-support ratings from 22 researchers, measuring tool experience rather than expert novelty labels.

Feasibility and proposal feedback. AI Idea Bench also derives overall and per-step feasibility proxies from the normalized, time-weighted citation impact of papers linked to decomposed idea steps. The authors interpret the per-step score as methodological soundness. However, citation impact does not test whether a retrieved method applies to the present step or whether steps are compatible; independent calibration against step-feasibility judgments is not reported. These scores directly measure the influence of associated literature.

ScholarEval and InnoEval seek richer assessments by producing evidence-bearing feedback (Moussa et al., 2026; Qiao et al., 2026). ScholarEval separates soundness and contribution analysis, evaluates coverage of 1,076 expert-checked feedback requirements, and studies report usefulness with 18 researchers. InnoEval combines papers, web resources, and code with multiple simulated reviewer perspectives; it evaluates historical decision labels, rankings, and expert dimension scores.

Synthesis. Ideation evaluation is most interpretable when novelty, feasibility, and feedback usefulness remain distinct. Blinding and style normalization control presentation; links to prior literature expose the basis of novelty judgments; expert-feedback requirements make coverage observable. These designs support interpretation without requiring agreement with a single published idea.

The unresolved feasibility question is whether influential supporting literature establishes that the

current plan is implementable. AI Idea Bench's citation-derived proxies provide evidence about the associated papers. Its per-step methodological-soundness interpretation lacks calibration against method applicability and step compatibility. Such judgments would test the proposed connection between its citation-derived proxies and feasibility. Practitioners could combine literature-grounded novelty assessment with separately calibrated feasibility judgments and feedback-usefulness studies. Expert comparisons should also report generation and selection budgets, which determine the proposals reaching evaluators.

3.3 Executable Research Workflow and Scientific Coding Evaluation

Executable workflow evaluation covers scientific calculations, data analysis, and machine-learning research engineering. Systems receive tasks and computational resources and return code, predictions, or analysis outputs. Running these artifacts enables direct checks, but the checks differ: successful execution, correct scientific outputs, defensible analysis decisions, and improved experimental performance are distinct targets.

Scientific code and analysis. SciCode uses scientist-curated problems with numerical and domain-specific tests (Tian et al., 2024). Its 65-problem test set contains 288 subproblems, allowing local implementations and their integration to be compared. Claude 3.5 Sonnet solves 26.0% of subproblems but 4.6% of main problems in the standard setting; supplied scientific background raises these to 35.4% and 12.3%. These conditions help diagnose knowledge and integration requirements. Because experts provide the decomposition, the results test implementation along that workflow rather than the system's ability to invent the decomposition.

ScienceAgentBench moves to self-contained Python programs for 102 data-driven scientific tasks (Chen et al., 2025d). It separates valid execution from task-specific success, using numerical checks, prediction thresholds, and, for some visual outputs, model scoring. For Claude's self-debugging configuration without expert knowledge, valid execution reaches 92.2%, but task success is 32.4%, under best-of-three selection. BLADE instead evaluates choices among defensible analyses: conceptual variables, data transformations, and statistical models are matched to expert reference decisions (Gu et al., 2025). Its finding that executable analyses can still cover few reference decision combinations illustrates what a run/no-run check misses. DiscoveryBench scores final hypotheses by matching their conditions, variables, and relationships (Majumder et al., 2024). Its no-data baseline is diagnostic: Llama-3 scores 11.5 without data versus 12.1 with CodeGen. This identifies a contribution from question and metadata information; the matching score alone does not isolate data verification.

Machine-learning engineering and research methods. MLE-bench packages 75 Kaggle competitions as offline tasks (Chan et al., 2025). Given a description and data, agents train models and submit predictions to a local hidden-test grader. Historical private-leaderboard thresholds translate task scores into medal attainment. Under AIDE, a 24-hour budget, and one A10 GPU, o1-preview attains a medal threshold in 16.9% of attempts; allowing eight attempts raises the at-least-one-success rate to 34.1%. Runtime and hardware specify conditions, while contamination checks audit the evidence. Because many tests are resplit from original training data, historical medals are calibrated references: the authors check example solutions and acknowledge that reconstructed and original competitions are not exactly equivalent.

RE-Bench supplies a complementary human comparison: 61 experts make 71 eight-hour attempts in seven research-engineering environments (Wijk et al., 2025). Relative performance changes with the budget. The reported 32-hour human result selects the best of four independent eight-hour attempts, rather than observing one continuous 32-hour session. This design makes time and attempt allocation explicit parts of the comparison.

MLRC-Bench connects proposed methods to implemented results in seven research competitions (Zhang et al., 2025b). Agents modify a protected repository, select a snapshot on development performance, and undergo hidden-test evaluation. Scores normalize improvement relative to baseline and human solutions and can be negative. Its auxiliary o1 novelty ratings have Spearman correlation -0.06 with measured effectiveness. Under this protocol, perceived innovativeness is a poor predictor of competition performance; effectiveness is not itself a novelty label, so the correlation does not establish that novelty judgments are inaccurate.

Synthesis. Executable evaluation gains diagnostic value by separating running code, obtaining correct results, and making appropriate analysis decisions. SciCode's local and integrated tests and ScienceAgentBench's execution and success measures locate different failures. BLADE checks analysis decisions against expert alternatives. DiscoveryBench's no-data control shows why performance must also be examined against information available without execution. Figure 1 distinguishes task checks (L2), direct analysis-decision evaluation (L3), and RE-Bench's controlled expert comparison (L4).

For research engineering, hidden-test scores provide an endpoint, while attempt selection, feedback, and compute define its conditions. MLE-bench and RE-Bench show why best-of-several outcomes and single-attempt success answer different operational questions. Comparisons should preserve both task success and resource cost. Decision-level checks add evidence about how an analysis was conducted. Their role depends on whether the intended use is output verification, workflow diagnosis, or performance comparison under a shared budget.

3.4 Academic Writing and Scholarly Communication Evaluation

Scientific writing evaluation asks whether a system expresses research accurately, uses sources appropriately, and communicates through a particular medium. These questions require different observations: reconstructing a citation paragraph, checking a reference, and reading a poster are distinct tasks. Their shared concern is how information survives selection, attribution, and presentation.

Source-supported writing. CiteBench standardizes citation-sentence and paragraph generation from supplied papers and citing context (Funkquist et al., 2023). It combines reference similarity with citation-intent and discourse distributions, and human ratings of readability and consistency. In its 150-output human study, average consistency remained below 3/5 across models, despite readable outputs. Reference reconstruction also depends on input sufficiency: the authors document targets containing facts absent from the permitted input, although the affected proportion is unquantified. Thus, some reconstruction errors cannot be attributed solely to the generator. At introduction level, SciIG separately evaluates reference-set recovery, content, and narrative quality from supplied metadata (Garg et al., 2026). The results show that its keyword-overlap measures and human factuality judgments have weak alignment, illustrating why metric names alone do not establish what they measure.

Task manipulation helps explain these differences. Across two models, six prompts, and six input configurations, Sahinuç et al. (2024) vary abstracts, citation intent, and example citations. Their exploratory 30-example human study finds that automatic measures provide only partial agreement with atomic-fact coverage. Reference-derived intent is an explicit experimental input, allowing the study to examine how task definition changes output. GREP instead evaluates related-work sections against hard constraints, such as source support, and soft preferences, such as citation emphasis and positioning (Şahinuç et al., 2026a). Its protocol includes 16 NLP/CV experts and criterion-specific comparisons. An initial 20% faithfulness pass rate for o3-mini means that every citation passed in 20% of sections, not that only 20% of citations were faithful. Reported citation-format parsing failures also enter this score, so it does not isolate semantic reasoning failures.

Reference and citation verification. Bibliographic validity and contextual support need different evidence. BibAgent retrieves source passages to diagnose miscitation; its paywall setting instead uses downstream citing papers as indirect evidence and can abstain (Li et al., 2026c). MisciteBench contains 6,350 expert-checked synthetic miscitations, all erroneous. Its Acc-pass@3 requires a correct validity label and an accepted explanation in at least one of three attempts; it therefore does not estimate false alarms on naturally correct citations. CiteAudit tests existence and metadata consistency on separate generated and real-world collections containing valid and invalid references (Shi et al., 2026). Its reported precision falls from .947 to .749 while recall remains .971 and .951, respectively. Separate confusion matrices expose differences that a single accuracy or synthetic-data result would conceal.

Visual communication. Paper2Poster's PaperQuiz asks six vision-language models to answer source-paper questions using only a poster (Pang et al., 2025). This measures information accessibility alongside visual and textual quality; raw answer accuracy is separate from its length-adjusted score. Human calibration

involved one PhD participant and five papers, limiting evidence for a general human-reader proxy. PaperBanana instead compares method diagrams on faithfulness, conciseness, readability, and aesthetics, with a hierarchical overall judgment (Zhu et al., 2026a). Separate 50-example studies calibrate the judge and collect blind preferences. Its overall improvement coexists with faithfulness below the reference baseline.

Synthesis. Writing evaluation becomes actionable when criteria identify revision needs. GREP separates unsupported content from unwanted emphasis or presentation choices. Citation-generation findings show that reference-based errors must be interpreted against the writer's inputs; separating parsing from semantic verification would likewise distinguish formatting failures from support failures. These distinctions help turn an overall writing score into a diagnosis that can guide revision.

Readability, faithfulness, and information access can diverge, as SciIG's weak metric–human alignment and PaperBanana's differing overall and faithfulness results illustrate. Reference verification additionally requires detecting errors without false alarms on valid references. Assessment should therefore combine criteria suited to the artifact and its readers: source support and author constraints for prose, and fidelity and information accessibility for visual summaries. Human calibration should test the particular reader judgment represented by each automated measure.

3.5 Automatic Peer Review Evaluation

A useful scientific review connects criticism to manuscript evidence and gives its recipient something interpretable or actionable. Evaluation accordingly examines review content, the validity of specific criticisms, and effects within the review process. The central design choice is which property the available reference or observation can check.

Review content and evaluator calibration. ReviewEval decomposes quality into alignment with human reviews, factual accuracy, analytical depth, constructiveness, and guideline adherence (Garg et al., 2025). Its factuality pipeline converts comments into questions, retrieves manuscript passages, generates a rebuttal, and checks support. Five experts also compare anonymous reviews on 30 papers, but their four preference dimensions do not supply independent labels for that question-level verifier. Consequently, high automatic support rates have different evidential backing from the reported human preferences. Peerispect makes verification explicit through supported, partially supported, contradicted, and undetermined labels (Ghorbanpour et al., 2026). Its constructed manuscript-claim set contains only supported claims; the separate real-review set supplies the multi-class test. GEM addresses another evaluator question: how much information a review shares with independent peer references when no unique gold review exists (Xu et al., 2025a). Teacher-score correlations and ICLR review-degradation tests are separate experiments. GEM is sensitive to information removal; unlike GPT-4o LMExaminer, its scores do not significantly increase after meaningless expansion in those tests. However, it does not outperform the LLM judge on every calibration measure.

Criticism and error localization. Natural weaknesses and controlled errors offer complementary reference construction strategies. CLAIMCHECK links expert-selected weaknesses to manuscript claims and separates claim association, weakness labeling/editing, and claim verification (Ou et al., 2025). Labeling/editing supplies the weakness; verification does not. The performance difference between these settings makes the input distinction essential. FLAWS constructs claim-invalidating edits and evaluates ranked error locations in 713 paper–error pairs (Xi et al., 2025). Its top-k score uses candidate depth within one output, with length restrictions on proposed spans. Human inspection validates 57 of 58 sampled errors, while a separate comparison checks automatic detection judgments. This provides evidence about both the constructed problems and their scorer.

SPOT instead contains 91 author-acknowledged errors in 83 historical manuscripts (Son et al., 2025). Models receive interleaved text and images; their descriptions are matched to known errors, and repeated runs support pass@K estimates. However, expert case analysis identifies a genuine unit error outside the annotations. Since unmatched findings count as false positives, matching precision cannot directly estimate the factual reliability of every criticism. PRISMM-Bench further isolates cross-modal inconsistencies through identification, remedy selection, and visual matching (Selch et al., 2026). Focused, page, and document contexts distinguish local recognition from broader inspection;

no-context and structured-option controls reveal how answer wording can support guessing.

Usefulness and workflow effects. RevUtil evaluates author-oriented actionability, grounding/specificity, verifiability, and helpfulness using 1,430 human-annotated review points (Sadallah et al., 2025). Its scorer sees the comment, not the paper: grounding means pointing to manuscript content, while verifiability concerns supplied reasons. A randomized ICLR deployment observes a different endpoint by providing feedback to reviewers after submission (Thakkar et al., 2025). Of 22,467 reviews assigned to feedback, 18,946 received it after gating. Review edits and discussion behavior are measured separately from blind preferences on 100 selected high-uptake revision pairs. The reported 89% preference for revisions concerns that subset.

Synthesis. Peer-review evaluation combines three views: complete-review quality, criticism correctness, and feedback effects. Multidimensional assessment, controlled or historical errors, and utility or deployment studies provide complementary evidence about content, detection, and use. Controlled errors make detection and localization testable, while historical errors retain problems encountered in real manuscripts. Human-annotated utility criteria and randomized deployment connect content to recipient judgments and workflow behavior. No single reference set supplies all three.

Calibration must match the scored property. A preference study assesses perceived review quality; an automated factuality checker needs independent judgments of manuscript support. ReviewEval illustrates this distinction, while FLAWS separately validates constructed errors and detection judgments. SPOT's valid finding outside its annotations shows why unmatched criticism needs adjudication before matching precision can represent factual reliability. Workflow studies add practical effects, but conclusions remain tied to feedback assignment, receipt, and the preference-study subset. Combining these checks would evaluate useful, well-grounded reviews while keeping content assessment, factual verification, and workflow effects separately interpretable.

3.6 End-to-End Automatic Scientific Research System Evaluation

End-to-end research evaluation follows a connected sequence of research decisions and outputs. The sequence depends on the task: some systems start with code templates, others with a research question or dataset. A multi-task suite may include such sequences alongside isolated tasks. Understanding the starting resources and final success criterion is therefore necessary before comparing overall scores.

Systems and evaluation environments. The AI Scientist connects ideation, experiments, figures, writing, and automated review within lightweight ML templates (Lu et al., 2024). Its counts of proposed ideas, completed experiments, and compiled papers expose different failure points. DeepScientist studies sustained exploration from reproduced baselines, combining task-native improvements, implementation records, and expert assessments (Weng et al., 2025b). Its reported results include human verification and substantial compute. These examples motivate reusable protocols while retaining each system's starting points and verification arrangements.

AstaBench provides 11 benchmarks spanning literature, code, analysis, and end-to-end discovery, with tool and cost metadata (Bragg et al., 2026). The suite's aggregate is not a single connected research run. Its E2E-Bench and Hard variant evaluate reports, code, and artifacts against required research steps. A step passes only when evidence does not conflict and at least one source supports completion. The original study's Table 21 reports that 16% of rubric-criterion assessments appeared satisfied in the paper but were contradicted by code or artifacts. Separately, a human spot check judged 92% of 50 development-set rubric scores to be correct. Mean step scores and all-required-step completion are reported separately; the non-executing Faker control receives partial step credit but completes no whole task. Cross-artifact checks thus identify inconsistencies that report-only evaluation would miss.

Research artifacts and executable outcomes. MLR-Bench uses workshop-derived open research questions and phase-specific rubrics (Chen et al., 2025b). Its 201 tasks support idea and proposal evaluation, whereas experiments, writing, and end-to-end runs mainly use ten selected tasks. MLR-Judge averages two model assessments and inspects execution records as well as papers. The authors report unverified synthetic or placeholder-data results in eight of ten Claude Code tasks, a concrete reason to inspect execution. Ten experts

provide calibration, but non-significant differences between human–model and human–human score discrepancies do not establish statistical equivalence.

ResearchGym supplies runnable repositories with source methods removed, retained baselines, and task-native graders (Garikaparthi et al., 2026). Under the initial 12-hour, single-A100 and $10 API budget, its main agent improves on the strongest visible baseline in one of 15 runs; mean subtask completion is 26.5%. Reporting completion, improvement, and best@3 separately distinguishes reliable execution from occasional gains. AIRS-Bench supplies data and evaluation scripts without initial solution code, accepting prediction files for hidden-label scoring (Lupidi et al., 2026). Its 24-hour, single-H200 repeated runs offer a contrasting starting condition, so results should be compared through both task design and resource settings.

Rediscovery and stage diagnosis. FIRE-Bench asks agents to recover established empirical findings while withholding the original implementation and conclusions (Wang et al., 2026b). For example, a question about information position can target the finding that relevant evidence is used better at the beginning or end of context than in the middle. The study runs 30 core and ten cross-domain problems, not the entire 100-paper source pool. Atomic-claim precision, recall, and F1 compare conclusions with reference findings; trajectory inspection attributes errors to planning, implementation, execution, and conclusion formation. Its added human check validates claim extraction, not the entire subsequent entailment scorer.

Synthesis. End-to-end evaluation needs explicit starting resources, required stages, and success criteria. Supplied templates, runnable baselines, and data-only starting points impose different research demands. Reporting stage and whole-task completion identifies where connected runs stop; hidden-test performance or reference-finding recovery measures their outcomes. Comparisons should retain these distinctions rather than assume equivalent starting points.

Cross-artifact verification makes completion judgments more diagnostic. AstaBench's report–artifact disagreements and MLR-Bench's placeholder-data cases show that code and experimental records can change the assessment of an apparently completed report. Process inspection tests whether claimed steps occurred and helps attribute failures; its findings should remain distinct from primary outcome scores, as with rediscovery matching and trajectory analysis. Evaluations should report starting assistance, verified completion, task-native outcomes, and resource use together, retaining successful runs and failures rather than only an aggregate artifact score.

4 Toward Credible Evaluation Infrastructure

The comparisons motivate task-dependent reporting (Table 2): input-sufficiency issues in CiteBench, the need for feasibility-proxy calibration in AI Idea Bench, problem and scorer checks in FLAWS, and report–artifact verification in AstaBench.

Comparability and calibration. Resource conditions and repeated attempts shape results: MLE-bench varies attempt budgets, and RE-Bench compares agents and experts at stated time budgets. Long-software-task research provides background on task duration and success measurement (Kwa et al., 2025). Evaluators also need checks: general LLM-judge studies document position and verbosity biases and examine possible self-enhancement bias (Zheng et al., 2023). Their occurrence and impact require verification in each scientific setting.

Reporting recommendations. Report the task and intended claim, inputs and data version, evaluator and scoring, sample and aggregation unit, and applicability conditions. Combine applicable modules in Table 2, such as error construction and matching for localization, or participants and outcomes for feedback utility. Public artifacts support inspection and reuse.

5 Conclusion

Across six targets, this survey compares evaluation designs, synthesizes findings and evidence gaps, and develops setting-specific reporting recommendations. These comparisons help readers select benchmarks, interpret results, and design subsequent evaluations.

Limitations

This review is selective and emphasizes automated research workflows in AI/ML with inspectable evaluation evidence. It does not exhaust the broader AI-for-Science or autonomous wet-lab literature.

Coverage is also uneven across disciplines because many available benchmarks concern computer science and machine learning. We synthesize reported studies and documented protocols rather than independently reproducing their experiments. The five claim distinctions guide interpretation; they are not a validated measurement scale or an exhaustive taxonomy of scientific value.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST, MYRG-GRG2026-00018-FIC, MYRG-CRG2026-00007-FIC), the Dr. Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. Surveygen: Quality-aware scientific survey generation with large language models. *Preprint*, arXiv:2508.17647.

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, and 20 others. 2026. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. *Preprint*, arXiv:2510.21652.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2025. Mle-bench: Evaluating machine learning agents on machine learning engineering. *Preprint*, arXiv:2410.07095.

Yifan Chang, Yukang Feng, Jianwen Sun, Jiaxin Ai, Chuanhao Li, S. Kevin Zhou, and Kaipeng Zhang. 2025. Sridbench: Benchmark of scientific research illustration drawing of image generation model. *Preprint*, arXiv:2505.22126.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *J Med Internet Res*, 26:e53164.

Guanhua Chen, Chuyue Huang, Yutong Yao, Shudong Liu, Xueqing Song, Lidia S. Chao, and Derek F. Wong. 2026. From scenes to elements: Multi-granularity evidence retrieval for verifiable multimodal RAG. *CoRR*, abs/2605.15019.

Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, and Derek F. Wong. 2025a. SGIC: A self-guided iterative calibration framework for RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 28357–28370. Association for Computational Linguistics.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. Scixgen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, volume EMNLP 2021 of *Findings of ACL*, pages 1483–1492. Association for Computational Linguistics.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. 2025b. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Preprint*, arXiv:2505.19955.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025c. Ai4research: A survey of artificial intelligence for scientific research. *Preprint*, arXiv:2507.01903.

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. 2025d. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *Preprint*, arXiv:2410.05080.

Martin Docekal, Martin Fajcik, and Pavel Smrz. 2026. Oarelatedwork: A large-scale dataset of related work sections with full-texts from open access sources. *Preprint*, arXiv:2405.01930.

Nils Dycke and Iryna Gurevych. 2026. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Preprint*, arXiv:2508.21422.

Tao Feng, Yihang Sun, and Jiaxuan You. 2025. Graph-eval: A lightweight graph-based llm framework for idea evaluation. *Preprint*, arXiv:2503.12600.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2023. Citebench: A benchmark for scientific citation text generation. *Preprint*, arXiv:2212.09577.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Mmreview: A multidisciplinary and multimodal benchmark for llm-based peer review automation. *Preprint*, arXiv:2508.14146.

Krishna Garg, Firoz Shaik, Sambaran Bandyopadhyay, and Cornelia Caragea. 2026. Towards ai-assisted research writing: Benchmarking llms for ai/ml introduction generation. *Preprint*, arXiv:2508.14273.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. 2025. Revieweval: An evaluation framework for ai-generated reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20542–20564. Association for Computational Linguistics.

Aniketh Garikaparthi, Manasi Patwardhan, and Arman Cohan. 2026. Researchgym: Evaluating language model agents on real-world ai research. *Preprint*, arXiv:2602.15112.

Ali Ghorbanpour, Soroush Sadeghian, Alireza Daghighfarsoodeh, Sajad Ebrahimi, Negar Arabzadeh, Seyed Mohammad Hosseini, and Ebrahim Bagheri. 2026. Peerispect: Claim verification in scientific peer reviews. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5161–5165. ACM.

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. 2025. Blade: Benchmarking language model agents for data-driven science. *Preprint*, arXiv:2408.09667.

Beichen Guo, Zhiyuan Wen, Yu Yang, Peng Gao, Ruosong Yang, and Jiaxing Shen. 2025. Sgsimeval: A comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems. *Preprint*, arXiv:2508.11310.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. 2024. Ideabench: Benchmarking large language models for research idea generation. *Preprint*, arXiv:2411.02429.

Hui Huang, Julien Velcin, and Yacine Kessaci. 2025. Towards multi-document question answering in scientific literature: Pipeline, dataset, and evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 10867–10881. Association for Computational Linguistics.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation. *Preprint*, arXiv:2310.03302.

Siyuan Huang, Yifan Zhou, Yutong Gao, Zi Yin, Juyang Bai, Xinxin Liu, Rama Chellappa, Chun Pong Lau, Cheng Peng, Sayan Nag, and Shraman Pramanick. 2026. Scifig: Towards automating editable figure generation for scientific papers. *Preprint*, arXiv:2601.04390.

Jena D. Hwang, Varsha Kishore, Amanpreet Singh, Dany Haddad, Aakanksha Naik, Malachi Hamada, Jonathan Bragg, Mike D'Arcy, Daniel S. Weld, Lucy Lu Wang, Doug Downey, and Sergey Feldman. 2026. Deep research, shallow evaluation: A case study in meta-evaluation for long-form qa benchmarks. *Preprint*, arXiv:2603.06942.

Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *Preprint*, arXiv:2502.16069.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2025. Can large language models unlock novel scientific research ideas? *Preprint*, arXiv:2409.06185.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, and 6 others. 2025. Measuring AI ability to complete long software tasks. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Avishek Lahiri, Yufang Hou, and Debarshi Kumar Sanyal. 2025. Taxoalign: Scholarly taxonomy generation using language models. *Preprint*, arXiv:2510.17263.

Bowen Li, Haochen Ma, Yuxin Wang, Jie Yang, Yining Zheng, Xinchi Chen, Xuanjing Huang, and Xipeng Qiu. 2026a. Beyond rating: A comprehensive evaluation and benchmark for ai reviews. *Preprint*, arXiv:2604.19502.

Mingzhe Li, Zhiqiang Lin, and Shiqing Ma. 2026b. Source or it didn't happen: A multi-agent framework for citation hallucination detection. *Preprint*, arXiv:2605.08583.

Peiran Li, Fangzhou Lin, Shuo Xing, Xiang Zheng, Xi Hong, Siyuan Yang, Jiashuo Sun, Zhengzhong Tu, and Chaoqun Ni. 2026c. Bibagent: An agentic framework for traceable miscitation detection in scientific literature. *Preprint*, arXiv:2601.16993.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *Preprint*, arXiv:2310.01783.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *Preprint*, arXiv:2501.04306.

Alisia Lupidi, Bhavul Gauri, Thomas Simon Foster, Bassel Al Omari, Despoina Magka, Alberto Pepe, Alexis Audran-Reiss, Muna Aghamelu, Nicolas Baldwin, Lucia Cipolina-Kun, Jean-Christophe Gagnon-Audet, Chee Hau Leow, Sandra Lefdal, Hossam Mossalam, Abhinav Moudgil, Saba Nazir, Emanuel Tewolde, Isabel Urrego, Jordi Armengol Estape, and 18 others. 2026. Airs-bench: a suite of tasks for frontier ai research science agents. *Preprint*, arXiv:2602.06855.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *Preprint*, arXiv:2407.01725.

Hanane Nour Moussa, Patrick Queiroz Da Silva, Daniel Adu-Ampratwum, Alyson East, Zitong Lu, Nikki Puccetti, Mingyi Xue, Huan Sun, Bodhisattwa Prasad Majumder, and Sachin Kumar. 2026. Scholareval: Research idea evaluation grounded in literature. *Preprint*, arXiv:2510.16234.

MZ Naser. 2026. How llms cite and why it matters: A cross-model audit of reference fabrication in ai-assisted academic writing and methods to detect phantom citations. *Preprint*, arXiv:2603.03299.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. 2025. Mlgym: A new framework and benchmark for advancing ai research agents. *Preprint*, arXiv:2502.14499.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S. Weld, Joseph Chee Chang, and Kyle Lo. 2024. Arxivdigestables: Synthesizing scientific literature into tables using language models. *Preprint*, arXiv:2410.22360.

Kazuya Nishimura, Kuniaki Saito, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Toward structured related work generation with novelty statements. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 38–57, Bangkok, Thailand. Association for Computational Linguistics.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. 2025. CLAIMCHECK: how grounded are LLM critiques of scientific papers? In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 21712–21735. Association for Computational Linguistics.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2poster: Towards multimodal poster automation from scientific papers. *Preprint*, arXiv:2505.21497.

Shuofei Qiao, Yunxiang Wei, Xuehai Wang, Bin Wu, Boyang Xue, Ningyu Zhang, Hossein A. Rahmani, Yanshan Wang, Qiang Zhang, Keyan Ding, Jeff Z. Pan, Huajun Chen, and Emine Yilmaz. 2026. Innoeval: On research idea evaluation as a knowledge-grounded, multi-perspective reasoning problem. *Preprint*, arXiv:2602.14367.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. Ai idea bench 2025: Ai research idea generation benchmark. *Preprint*, arXiv:2504.14191.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2026. Human-llm compound system for scientific ideation through facet recombination and novelty evaluation. *Preprint*, arXiv:2409.14634.

Delip Rao and Chris Callison-Burch. 2026. Bibtex citation errors in scientific publishing agents: Evaluation and mitigation. *Preprint*, arXiv:2604.03159.

Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. 2026. Towards scientific intelligence: A survey of llm-based scientific agents. *Preprint*, arXiv:2503.24047.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2026. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Preprint*, arXiv:2412.17596.

Hyun Ryu, Doohyuk Jang, Hyemin S. Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, Haechan Kim, Jina Kim, Joowon Kim, Yoonjeon Kim, Kwanhyung Lee, Chanjae Park, Heecheol

Yun, Gregor Betz, and Eunho Yang. 2026. Reviewscore: Misinformed peer review detection with large language models. *Preprint*, arXiv:2509.21679.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 28991–29021. Association for Computational Linguistics.

Furkan Sahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with llms: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4832–4855. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in acl conferences. *Preprint*, arXiv:2601.18724.

Lukas Selch, Yufang Hou, M. Jehanzeb Mirza, Sivan Doveh, James Glass, Rogerio Feris, and Wei Lin. 2026. Prismm-bench: A benchmark of peer-review grounded multimodal inconsistencies. *Preprint*, arXiv:2510.16505.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing, SDP@ACL 2025, Venice, Austria, July 31, 2025*, pages 96–113. Association for Computational Linguistics.

Yuexi Shen, Minqian Liu, Dawei Zhou, and Lifu Huang. 2026. Navigating ideation space: Decomposed conceptual representations for positioning scientific ideas. *Preprint*, arXiv:2601.08901.

Kaiwen Shi, Weixiang Sun, Zheyuan Zhang, Lichao Sun, Nitesh V. Chawla, and Yanfang Ye. 2026. Citeaudit: You cited it, but did you read it? a benchmark for verifying scientific references in the llm era. *Preprint*, arXiv:2602.23452.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Mind the blind spots: A focus-level evaluation framework for LLM reviews. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 35630–35656. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *Preprint*, arXiv:2409.13740.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. 2025. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research. *Preprint*, arXiv:2505.11855.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. 2026. Surge: A benchmark and evaluation framework for scientific survey generation. *Preprint*, arXiv:2508.15658.

Maojun Sun, Yifei Xie, Yue Wu, Ruijian Han, Binyan Jiang, Defeng Sun, Yancheng Yuan, and Jian Huang. 2026. Dsaeval: Evaluating data science agents on a wide range of real-world data science problems. *Preprint*, arXiv:2601.13591.

Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Wenhao Huang, Ge Zhang, Jian Yang, and Zhoujun Li. 2025a. P2p: Automated paper-to-poster generation and fine-grained benchmark. *Preprint*, arXiv:2505.17104.

Zhaojun Sun, Xuzhou Zhu, Xuanhe Zhou, Xin Tong, Shuo Wang, Jie Fu, Guoliang Li, Zhiyuan Liu, and Fan Wu. 2025b. Surveybench: Can llm(-agents) write academic surveys that align with reader needs? *Preprint*, arXiv:2510.03120.

Xuemei Tang, Xufeng Duan, and Zhenguang G. Cai. 2025. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 1602–1617. Association for Computational Linguistics.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *Preprint*, arXiv:2504.09737.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, and 11 others. 2024. Scicode: A research coding benchmark curated by scientists. *Preprint*, arXiv:2407.13168.

William H. Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1):14045.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 279–299. Association for Computational Linguistics.

Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, and Daniel Khashabi. 2026a. arxiv2table: Toward realistic benchmarking and evaluation for llm-based literature-review table generation. *Preprint*, arXiv:2504.10284.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. *Preprint*, arXiv:2406.10252.

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. 2026b. Fire-bench: Evaluating ai agents on the rediscovery of scientific insights. *Preprint*, arXiv:2602.02905.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, and 8 others. 2025. From ai for science to agentic science: A survey on autonomous scientific discovery. *Preprint*, arXiv:2508.14111.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025a. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2025b. Deepscientist: Advancing frontier-pushing scientific findings progressively. *Preprint*, arXiv:2509.26603.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, and 4 others. 2025. Rebench: Evaluating frontier ai r&d capabilities of language model agents against human experts. *Preprint*, arXiv:2411.15114.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Xiangkun Hu, Pengrui Lu, Xiaojie Cai, Lyumanshan Ye, Wenhong Zhu, Yang Xiao, and Pengfei Liu. 2025. Innovatorbench: Evaluating agents' ability to conduct innovative llm research. *Preprint*, arXiv:2510.27598.

Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. 2025. Flaws: A benchmark for error identification and localization in scientific papers. *Preprint*, arXiv:2511.21843.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, Jian Wu, and Yue Zhang. 2025. How far are ai scientists from changing the world? *Preprint*, arXiv:2507.23276.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. 2025a. Benchmarking llms' judgments with no gold standard. *Preprint*, arXiv:2411.07127.

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025b. Can llms identify critical limitations within scientific research? a systematic evaluation on ai research papers. *Preprint*, arXiv:2507.02694.

Zuyao Xu, Yuqi Qiu, Lu Sun, Fasheng Miao, Fubin Wu, Xiang Li, Xinyi Wang, Haozhe Lu, Zhengze Zhang, Yuxin Hu, Jialu Li, Luo Jin, Feng Zhang, Rui Luo, Xinran Liu, Yingxian Li, and Jiaji Liu. 2026. Ghostcite: A large-scale analysis of citation validity in the age of large language models. *Preprint*, arXiv:2602.06718.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *Preprint*, arXiv:2503.04629.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Preprint*, arXiv:2509.00496.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. *Preprint*, arXiv:2305.14647.

Wenzheng Zeng, Mingyu Ouyang, Langyuan Cui, and Hwee Tou Ng. 2025. Slidetailor: Personalized presentation slide generation for scientific papers. *Preprint*, arXiv:2512.20292.

Guo-Biao Zhang, Xian-Ling Mao, Ding-Yuan Liu, Da-Yi Wu, Tian Lan, Huihui Li, and Heyan Huang. 2026a. Deepsurvey-bench: Evaluating academic value of automatically generated scientific surveys. *Preprint*, arXiv:2601.15307.

Ming Zhang, Jiabao Zhuang, Wenqing Jing, Kexin Tan, Ziyu Kong, Jingyi Deng, Yujiong Shen, Yuhui Wang, Zhenghao Xiang, Qiyuan Peng, Yuhang Zhao, Ning Luo, Renzhe Zheng, Jiahui Lin, Mingqi Wu, Long Ma, Shihan Dou, Maxm Pan, Tao Gui, and 2 others. 2026b. Can deep research agents retrieve and organize? evaluating the synthesis gap with expert taxonomies. *Preprint*, arXiv:2601.12369.

Tianmai M. Zhang and Neil F. Abernethy. 2024. Detecting reference errors in scientific literature with large language models. *Preprint*, arXiv:2411.06101.

Yanbo Zhang, Sumeer A. Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, Saso Dzeroski, Jesper Tegner, and Hector Zenil. 2025a. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1(1):14.

Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, and Kan Ren. 2024. Benchmarking data science agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5677–5700. Association for Computational Linguistics.

Yunxiang Zhang, Muhammad Khalifa, Shitanshu Bhushan, Grant D Murphy, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2025b. Mlrc-bench: Can language agents solve machine learning research challenges? *Preprint*, arXiv:2504.09702.

Jiahao Zhao, Shuaixing Zhang, Nan Xu, and Lei Wang. 2025. Surveyeval: Towards comprehensive evaluation of llm-generated academic surveys. *Preprint*, arXiv:2512.02763.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 17733–17750. Association for Computational Linguistics.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9340–9351. ELRA and ICCL.

Dawei Zhu, Rui Meng, Yale Song, Xiyu Wei, Sujian Li, Tomas Pfister, and Jinsung Yoon. 2026a. Paperbanana: Automating academic illustration for ai scientists. *Preprint*, arXiv:2601.23265.

Minjun Zhu, Zhen Lin, Yixuan Weng, Panzhong Lu, Qiujie Xie, Yifan Wei, Sifan Liu, Qiyao Sun, and Yue Zhang. 2026b. Autofigure: Generating and refining publication-ready scientific illustrations. *Preprint*, arXiv:2602.03828.

Qiran Zou, Hou Hei Lam, Wenhao Zhao, Yiming Tang, Tingting Chen, Samson Yu, Tianyi Zhang, Chang Liu, Xiangyang Ji, and Dianbo Liu. 2026. Fml-bench: Benchmarking machine learning agents for scientific research. *Preprint*, arXiv:2510.10472.

Furkan Şahinuç, Subhabrata Dutta, and Iryna Gurevych. 2026a. Expert preference-based evaluation of automated related work generation. *Preprint*, arXiv:2508.07955.

Furkan Şahinuç, Subhabrata Dutta, and Iryna Gurevych. 2026b. Reward modeling for scientific writing evaluation. *Preprint*, arXiv:2601.11374.

A Comparative Mapping and Reporting Recommendations

Table 1 summarizes selected applications of the claim-strength analysis. Table 2 records the task-dependent reporting recommendations discussed in Section 4.

Task / representative benchmark	Evidence, evaluator, and metric	Diagnostic information	Supported interpretation and conditions
Survey assessment / SurveyBench (Sun et al., 2025b)	Human reference surveys and generated text; LLM content scores and quizzes answered from retrieved passages	Compares content scores with question answerability	**L2.** Reference-relative scores and simulated information use on 20 topics; human-written references do not themselves validate agreement with reader judgments
Idea assessment / AI Idea Bench (Qiu et al., 2025)	Target papers and retrieved literature; LLM matching judgments and citation-impact feasibility proxies	Separates reference matching from a literature-impact proxy for feasibility	**L2.** Reference-conditioned idea matching; citation impact characterizes associated literature, and the feasibility proxy is not calibrated to step applicability or compatibility
ML engineering / MLE-bench (Chan et al., 2025)	Competition data and submissions; local grader, historical medal thresholds, and pass@k	Measures task completion and performance at stated attempt budgets	**L2.** ML-engineering task performance under the specified environment, resources, and attempt budget; historical medal thresholds provide a calibrated human-performance reference, while resplit test sets and reimplemented graders limit direct comparison with the original competitions
Related-work writing / GREP (Şahinuç et al., 2026a)	Source papers, author preferences, generated sections, and revisions; criterion-specific checks and expert pairwise judgments	Separates source-support constraints from preferences about length, emphasis, and positioning; tests feedback-driven revision	**L4.** Expert-aligned assessment of related-work writing in the tested NLP/CV settings, alongside constraint checks; whole-section faithfulness scores include citation-parsing failures
Generated-review assessment / ReviewEval (Garg et al., 2025)	Manuscripts, reference reviews, venue rules, and generated reviews; alignment, factual-support, depth, constructiveness, and guideline scores	Compares dimensions of review quality; checks comments through manuscript retrieval and generated rebuttals	**L2.** Multidimensional review assessment under manuscript and reference checks; supplementary expert preferences on 30 papers do not independently calibrate the factual-support verifier
Integrated execution / AstaBench (E2E) (Bragg et al., 2026)	Reports, code, and artifacts; rubric-based step scores and whole-task completion	Cross-checks reported work against execution artifacts	**L3.** Completion of specified research tasks with research-step verification; artifact checks expose discrepancies missed by report-only inspection

Table 1: Selected applications of the claim-strength analysis in Figure 1. Labels L2–L4 summarize the core evaluation design described in each row. Scope conditions and identified support gaps are distinguished in the text.

Evaluation target	When applicable	Evidence and checks	Protocol details to report	Purpose
Literature synthesis	Source coverage and organization	Retrieved and supplied papers, reference structures, generated synthesis, and source-support checks	Retrieval cutoff and depth, reference construction, valid alternatives, and scoring denominators	Locate losses in retrieval, selection, or synthesis; interpret reference coverage
Literature synthesis	Answerability or reader benefit	Questions, answer evidence, reader tasks, and reference or control conditions	Human versus simulated readers, evidence access, participant sample, and aggregation	Distinguish supported answerability from observed reader benefit
Research ideation	Novelty or feasibility assessment	Prior-work comparisons, decomposed proposals, expert labels, and proxy-calibration results	Literature cutoff, retrieval coverage, judgment criteria, and reference-label construction	Identify what novelty and feasibility judgments are supported by
Research ideation	Proposal feedback or user benefit	Feedback requirements, evidence-bearing reports, and researcher judgments or revisions	Expert selection, blinding, agreement, comparison conditions, and measured outcomes	Relate feedback coverage to perceived usefulness or observed improvement
Executable workflows	Code or task correctness	Submitted programs, output tests, reference solutions, and execution results	Inputs, supplied decomposition, tests, environment, and permitted feedback	Separate successful execution from task-specific correctness
Executable workflows	Analysis decisions or performance comparison	Checked transformations and model choices; task scores and human attempts where available	Tool access, runtime, compute, retries, selection rules, and reference decisions	Explain analysis choices and make resource-conditioned comparisons interpretable
Scholarly writing	Section generation and revision	Source texts, author constraints, generated drafts, and criterion-specific judgments	Input sufficiency, parsing rules, rubric, evaluator calibration, and revision rounds	Separate content support, preference satisfaction, and formatting failures
Scholarly writing	Citation verification or visual communication	Source passages and validity labels for citations; source content, visual artifacts, and reader checks for communication	For verification: access limits and unavailable-source handling; for communication: fidelity criteria and reader conditions	Make citation decisions traceable; assess preserved content and accessibility
Automatic peer review	Generated-review quality	Manuscript, generated review, aspect scores, and independent judgments for the assessed dimensions	Evaluator, rubric, model/prompt, blinding, agreement, and calibration for each dimension	Distinguish critique support, analytical depth, constructiveness, and preference
Automatic peer review	Error discovery or feedback use	Reference errors and matched findings for detection; assignments, revisions, and outcomes for feedback	For detection: error construction and unmatched findings; for use: participants, delivery, uptake, and controls	Interpret detection coverage separately from observed workflow effects
End-to-end research	Connected research execution	Stage outputs, code, logs, reports, and checked experimental artifacts	Starting resources, handoffs, execution checks, tools, time, compute, and retries	Verify reported work and attribute failures across stages
End-to-end research	Overall completion or suite aggregation	Required steps, final outcomes, subtask scores, and missing or failed runs	Connected-run versus suite design, weights, denominators, and resource allocation	Separate whole-task completion from average component performance

Table 2: Reporting recommendations organized by evaluation target and setting. Common fields are stated in Section 4. Requirements apply to the named setting; a cross-target protocol may use several rows.

B Search and Protocol Records

The following tables document protocols within the review's six targets, including the studies discussed in the main text. Rows distinguish task inputs, the evidence used to judge outputs, scoring, calibration or controls, and findings or applicability conditions. Multiple protocols from one work are distinguished when their evidence or scoring differs.

Search terms combined each target with evaluation-related variants. Literature synthesis queries included survey generation, cross-paper comparison, and survey evaluation; ideation queries included research idea evaluation and novelty assessment; workflow queries included scientific code, data analysis, and ML research agents. Writing queries included citation-text generation, citation support, and scientific communication artifacts; reviewing queries included grounded critique, error localization, and review utility; end-to-end queries included research agents, research environments, and scientific rediscovery. Backward and forward citation tracing connected these searches. The tables preserve relevant alternative designs rather than all neighboring application areas.

B.1 Scientific Literature Synthesis Protocols

The protocols below distinguish survey generation, structured synthesis, and question-driven information use. Separate rows identify different scoring tasks within a work; human reference artifacts and independent human judgments are recorded separately.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Survey generation and content assessment				
AutoSurvey: survey quality (Wang et al., 2024b)	Topic and retrieved papers; generated sections and citation passages.	Multiple LLMs score coverage, structure, relevance, and citation support.	Same writer/reference budget for naive retrieval baseline; three doctoral raters rank 20 surveys.	Citation recall covers already-cited sentences; ranking correlation with humans is 0.5429.
AutoSurvey: answerability (Wang et al., 2024b)	Survey passages support 50 multiple-choice questions from five topics.	LLM question-answering accuracy.	No material, naive retrieval survey, and option-specific retrieval conditions.	AutoSurvey improves accuracy by 9.2 points over no material, 2.4 over naive retrieval; simulated use.
SurGE (Su et al., 2026)	205 human surveys; shared title/abstract retrieval pool; generated references, outlines, and sections.	Reference recall; document/section/sentence citation judgments; structure/content scores.	Expert anonymous rankings calibrate structure/content; shared backbone for dedicated generators.	Citation support uses abstracts; ranking calibration does not cover citation judgments.
SurveyGen/QUAL-SG (Bao et al., 2025)	Topic-only, automatic retrieval, or supplied human references/outline; generated surveys.	Reference authenticity and overlap; key-point recall, text/structure similarity, LLM ratings.	120 surveys in four fields; three researchers compare 15 CS outputs with human originals.	Resource contains 4,205 surveys, not 4,205 evaluated outputs; supplied outlines change the task.
Tang et al.: bibliography (Tang et al., 2025)	Title/keywords of Annual Reviews articles; ten generated references per query.	Semantic Scholar metadata matching; authenticity precision and reference overlap.	Three raters check 100 references; automated accuracy 86%, $\kappa = 0.71$.	Matching allows partial metadata agreement; authenticity and overlap have different meanings.

Table 3: Detailed protocols for scientific literature synthesis.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Tang et al.: abstracts (Tang et al., 2025)	Title/keywords; generated and human abstracts of matched length.	Semantic similarity, overlap, and two textual-entailment judges.	Five-model comparison; content reference is the human abstract.	Published entailment formula tests generated-to-reference direction, not all extra generated assertions.
Tang et al.: short reviews (Tang et al., 2025)	Title, keywords, and human abstract; approximately 1,000-word review with ten citations.	Bibliographic checks, overlap, and recall of ten reference key points.	Same 1,105-review source collection; tasks are not a chained pipeline.	Human abstract is supplied; later GPT-5 supplement uses a different 2025 collection.
SurveyForge (Yan et al., 2025)	Topic, retrieved papers, and human outline examples; survey and bibliography.	Reference-set overlap; LLM outline/content scores.	20 doctoral researchers blind-compare 100 outputs with AutoSurvey; same-backbone comparisons.	Human outline preference is 74% against AutoSurvey, not against human surveys; reference formula is precision-like.
SurveyEval (Zhao et al., 2025)	38 topics; human and generated surveys; reference lists.	Human-anchored content rubric; reference-free outline rubric; bibliography precision/recall/F1.	Common title queries; reference-list evaluation only for two of seven systems.	Human text is an anchor, not independent human calibration; overlap does not identify hallucinated references.
SGSimEval (Guo et al., 2025)	80 human surveys and collected system outputs; outline/content/reference embeddings.	LLM ratings with similarity weighting toward full marks or measured human-text scores.	Alternative LLM pairwise comparisons; no direct expert-score calibration reported.	Some systems only cover CS topics; weighted scores describe unchanged outputs under another scale.
DeepSurvey-Bench: rubric (Zhang et al., 2026a)	178 selected human surveys; generated surveys; seven academic-value criteria.	Three-model ratings; human agreement and correlations with surface-quality scores.	Three human raters assess 40 generated surveys; separate reference-selection annotation.	Rubric and surface scores correlate only partly in this sample.
DeepSurvey-Bench: evidence audit (Zhang et al., 2026a)	400 claim–evidence units from 40 high-scoring surveys, checked against sources.	Human factual correctness, evidence alignment, and citation reliability.	Separate from the rubric-calibration protocol; selection is conditional on high scores.	Reported values are 94%, 91%, and 93%; these are not whole-corpus accuracy estimates.
DeepSurvey-Bench: feedback (Zhang et al., 2026a)	40 low-scoring surveys revised once with rubric feedback.	Same model ensemble scores before/after on seven criteria.	Topic, outline, scope, and reference set held fixed.	Higher scores are reported; independent human assessment of revision gains is not reported.
Comparison tables and field organization				
ArxivDIGESTables: columns (Newman et al., 2024)	Titles/abstracts and optional captions/context; generated comparison dimensions.	Semantic column alignment and schema recall; human usefulness, specificity, and insightfulness.	50-table alignment audit; generated column count fixed to reference count.	Alignment precision 70–85%, depending on partial-match treatment; unmatched columns can be useful.

Table 3 (continued): Detailed protocols for scientific literature synthesis.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
ArxivDIGESTables: cells (Newman et al., 2024)	Human schema and paper full text; generated cell values.	Automatic semantic matching; human full/partial/no-match labels.	30-table audit, two raters per value; context variants.	Extra context mainly improves partial matches; this is conditional filling, not end-to-end table accuracy.
arXiv2Table (Wang et al., 2026a)	Rewritten request, papers and annotated distractors; selected papers and comparison table.	Paper-selection precision/recall/F1; bidirectional questions on schema, cells, and cell pairs.	Human checks of 200 questions and 300 judgments per method; judge-swap controls.	Main tasks use candidate pools; novel content beyond the reference receives no credit by design.
TaxoBench (Zhang et al., 2026b)	Topic-only retrieval/organization or supplied expert paper set; hierarchy and paper assignments.	Paper overlap, leaf clustering, semantic tree/path distances, and tree statistics.	Random/depth controls; two independent human organizers on ten topics.	Path scores depend on labels/depth; ten-topic human comparison is not a 72-topic estimate.
TaxoAlign/CS-TAXOBENCH (Lahiri et al., 2025)	Topic and supplied reference papers; generated classification tree; extracted human trees.	Node recall, text similarity, size ratio, and model/human structure/content ratings.	Three raters on 20 trees; reference-extraction audit; additional conference test set.	References are supplied. Degree-sum ratio measures tree size, not parent–child alignment.
Question-driven synthesis and information use				
OpenScholar: ScholarQABench (Asai et al., 2024)	Scientific questions and open retrieval; cited answers, expert answers, and rubrics.	Automatic answer/citation scores; expert quality, usefulness, and pairwise preference.	108 questions, 16 experts; no self-question rating; 50-question length control.	Expert preference and citation correctness are separate; experimental corpus is peS2o v2.
ResearchQA (Yifei et al., 2025)	Survey-mined questions and requirements; long scholarly answers with citations.	LLM requirement coverage, pairwise preference, and tournament ranking.	31 doctoral-level annotators; validation in eight of 75 fields; length sensitivity on 225 questions.	Coverage is recall-like; incomplete citation requirements do not imply fabricated citations.
SurveyBench: content (Sun et al., 2025b)	20 topics, human references, generated surveys and outlines.	LLM outline/content rubrics; non-text-element density.	Reference versus no-reference scoring; human quality anchors fixed to full marks.	DeepResearch content mean 4.42/5; anchor-based scores are not independent expert ratings.
SurveyBench: quizzes (Sun et al., 2025b)	General/topic questions; passages retrieved from evaluated survey; cited answers.	LLM answer quality and evidence checks; unsupported answer scored zero.	Human-survey answering reference; topic-specific anchor fixed to 10.	Topic-specific scores 1.48–3.19/10 across four systems; measures simulated supported answerability.
Deep Research, Shallow Evaluation (Hwang et al., 2026)	Scientific reports; holistic preferences and criterion-specific expert annotations.	Agreement and rank correlation at system and instance levels.	200-question holistic study; separate near/deep-expert studies; 80 fixed-claim citation checks.	System agreement can conceal weaker instance agreement; expertise comparisons also change questions and system pool.
MDA-QA (Huang et al., 2025)	Citation-linked paper communities; 6,804 filtered questions requiring multiple sources.	Support-document recall/exact-match retrieval; answer similarity metrics.	Single-document answer screening; separate 30- and 200-question expert checks.	198/200 audited questions meet all criteria; automatic filtering is not a strict necessity proof for every item.

Table 3 (continued): Detailed protocols for scientific literature synthesis.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
PaperQA2/LitQA2 (Skarlinski et al., 2024)	248 full-text-evidence questions; retrieval traces and answers.	Answer accuracy/precision and supporting-DOI recall.	Nine doctoral researchers; three system runs; contaminated human answers excluded.	Answer precision 85.2%; accuracy 66.0%, not significantly above human accuracy.
PaperQA2/WikiCrow (Skarlinski et al., 2024)	Generated gene entries paired with existing Wikipedia entries; cited statements.	Human citation-support categories and precision.	375 statement evaluations by four experts, blinded to source.	Experts judge existing entries; they do not write matched new comparison articles.
PaperQA2: ContraCrow (Skarlinski et al., 2024)	Claims from 93 biology papers; retrieved contradictory evidence.	Graded contradiction judgments and human candidate validation.	100 high-scoring candidates audited; separate detection setting without system reasoning.	70% of sampled candidates validated; human validation uses supplied evidence, not unrestricted search.

Table 3 (continued): Detailed protocols for scientific literature synthesis.

B.2 Research Ideation and Proposal Protocols

Expert judgments, published-paper references, literature-relative novelty, and proposal feedback use different labels. The table separates assessment of ideas from assessment of interactive support or report usefulness.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Expert judgments and reference-driven comparisons				
Si et al. (Si et al., 2024)	Human, AI-selected, and human-reranked AI proposals; shared templates and seven NLP topics.	Blinded expert novelty, excitement, feasibility, expected effectiveness, and overall ratings.	104 distinct researchers; 49 proposals per condition; style normalization and multiple statistical units.	Higher AI novelty ratings; feasibility differences not significant at idea level. AI uses extensive sampling.
AI Idea Bench: target matching (Qiu et al., 2025)	3,495 target papers; inspiration papers/topic; generated motivation and plan.	LLM multiple-choice identification, idea/topic matching, and pairwise ranking.	Common generation backbone; two candidates with best-of-two aggregation; researchers screen inspiration papers.	Motivation and plan rankings differ; one baseline lacks plans. Target matching is reference recovery.
AI Idea Bench: literature proxies (Qiu et al., 2025)	Idea components linked to prior/contemporary papers and citation statistics.	Literature-distance novelty score; citation-impact feasibility and per-step scores.	Time weighting and citation normalization; no independent step-feasibility calibration reported.	Citation influence does not check method applicability or step compatibility; in the original study's Eq. (7), larger historical distance lowers the novelty score when other terms are fixed.
IdeaBench (Guo et al., 2024)	Biomedical target-paper references; three generated ideas mixed with target idea.	GPT-4o relative novelty/feasibility ranking; semantic similarity and overlap.	Five filtered references versus all references; blinded origins; fixed candidate count.	Insight Score is the fraction ranked above the target; authors report high scores for some incoherent outputs.
AP-FRI (Kumar et al., 2025)	1,250 papers with future-work text removed; generated directions and author-written references.	Idea matching and semantic diversity; expert novelty, relevance, feasibility/correctness.	Matcher checked on labeled pairs; author studies in CS and physics; retrieval comparison.	Author alignment does not exhaust valid directions; the original study's Eq. (1) sums over generated ideas, while the following text defines the denominator as the author's idea count. The five Claude novelty percentages in the original study's Section 5.3 sum to 110.41%.

Table 4: Detailed protocols for research ideation and proposal evaluation.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
LiveIdeaBench (Ruan et al., 2026)	Single scientific keywords; short ideas across 1,180 keywords in 22 disciplines.	Model-panel originality, feasibility, clarity, diversity; lower-quantile aggregate.	Self-judging prohibited; six differential-equations experts assess 22 ideas; dynamic judge selection.	Domain calibration uses all eligible judges, unlike main random panels; quantile is keyword-weighted.
GraphEval: historical labels (Feng et al., 2025)	Paper titles/abstracts, viewpoint graph, and known training labels.	Label propagation or graph network predicts historical decision classes; accuracy/macro-F1.	ICLR train/test split; alternative prompting/classifier baselines.	Supervised reference is a decision label; AI Researcher baselines use differing test scopes.
GraphEval: derived ideas (Feng et al., 2025)	Copied, replaced, or recombined viewpoints with time features.	Classification of synthetic derivative ideas.	80 constructed examples; ten used to create training negatives.	Reported detection improvement concerns constructed derivatives; the original study's Section 7.2 does not state whether the ten ideas used to construct training negatives are excluded from testing.
Literature-relative novelty and interactive support				
Literature-Grounded Novelty Assessment (Shahid et al., 2025)	Idea and retrieved top-ten papers; expert examples and binary labels with reasons.	Accuracy, precision/recall/F1, and Cohen's kappa.	35 training/32 test ideas; common papers for prompt comparisons; separate 58 non-novel cases.	Test accuracy 0.81, $\kappa = 0.59$; 58-case result is restricted to non-novel ideas.
Navigating Ideation Space (Shen et al., 2026)	Problem, method, and findings representations; component-linked prior papers and expert idea scores.	Retrieval ranking metrics; similarity-derived novelty correlation.	466 valid retrieval queries; 93 expert-scored ideas; retrieval-budget comparisons.	Spearman 0.384 versus GPT-5 Mini 0.385; does not establish superiority on all measures.
Scideator: user support (Radensky et al., 2026)	Seed papers and selected problem/method/evaluation facets; interaction logs and questionnaires.	Creativity Support Index, confidence, interviews, and paired tests.	22 researchers, two tools for 20 minutes each; same LLM, different tool configurations.	Median support 70.50 versus 61.00; not isolated facet-only effect.
Scideator: novelty module (Radensky et al., 2026)	Retrieved papers, generated ideas, expert novelty examples and labels.	Binary novelty accuracy/F1; known-non-novel retrieval ablations.	Shared evidence with Literature-Grounded Novelty Assessment; reported split sizes conflict.	The module table does not establish an independent replication. The original study's Section 2 reports 67 labeled items and 20 in-context examples per class, while the original study's Section 4.2.2 reports 43 held-out items; these counts do not specify a consistent disjoint split.

Table 4 (continued): Detailed protocols for research ideation and proposal evaluation.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
SciMON (Wang et al., 2024a)	Background, optional seed, and literature inspirations; generated idea sentence and revisions.	Expert helpfulness, technical detail, relative novelty; auxiliary text similarity.	Separate generation and iteration studies; 194 human-cleaned test cases; preliminary biomedical study.	Changed wording and increased novelty are separately judged; supplied seed/reference conditions are explicit.
Proposal reports and their evaluation				
ScholarEval: report assessment (Moussa et al., 2026)	Problem, method, and experiment plans; full-text soundness evidence and abstract-based contribution comparisons.	Coverage of 1,076 expert-checked requirements; report preferences; invalid-link rate.	117 ideas, six rubric validators; dated retrieval; module ablations.	Zero links failing specified HTTP checks does not establish absence of all citation-support errors.
ScholarEval: user study (Moussa et al., 2026)	Researchers' ideas and blinded ScholarEval/deep-research reports.	Useful citations and report-quality/refinement questionnaire scores; mixed-effects analysis.	18 experts, 46 evaluations; different ideas used with each system.	More useful references and higher refinement-related ratings in this study.
InnoEval (Qiao et al., 2026)	Timestamped ideas, paper/web/code evidence, and simulated reviewer perspectives.	Historical-label accuracy/F1, pair/group ranking, report preference, expert-score correlation.	Common DeepSeek-V3.2 backbone; 60 expert-rated ideas; knowledge-time separation.	Best-choice accuracy 65.12%, exact ranking 22.09%; distinct ranking endpoints.

Table 4 (continued): Detailed protocols for research ideation and proposal evaluation.

B.3 Executable Workflow and Scientific Coding Protocols

These protocols inspect scientific functions, analysis decisions, program outputs, or machine-learning improvements. Execution tests, model judgments, and process measurements remain separate; resource and feedback conditions specify how each result was obtained.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Scientific code and analysis decisions				
SciCode (Tian et al., 2024)	Scientist-defined functions and workflow; generated code, reference implementations, numerical/domain tests.	Subproblem and main-problem pass@1; main success requires steps and integration.	At least two domain reviewers per problem; optional background and gold-prior-step controls.	65 main/288 subproblems in test set. Decomposition is provided, not independently generated.
BLADE: decision selection (Gu et al., 2025)	Research question/data; 188 questions on most/least defensible analysis choices.	Multiple-choice accuracy.	Expert-derived reference decisions; deterministic generation setting.	Measures decision recognition; separate from implementing an analysis.
BLADE: analysis generation (Gu et al., 2025)	12 research questions/data sets; variables, transformation functions, and statistical-model functions.	Semantic, executed-value, and dataflow matching; precision, coverage@10, combined F1.	11 experts construct 536 decisions; two authors audit 615 evaluator outputs.	GPT-4o produces runnable nonempty analyses in 96% of cases, yet model/variable coverage stays below 13%.
DiscoveryBench (Majumder et al., 2024)	Real/synthetic data, metadata, and constrained question; generated hypothesis.	LLM-assisted matching of conditions, variables, and relationships, on a 0–100 scale.	No-data baseline; separate oracle-score-feedback condition; real test set has 239 tasks.	Llama-3 no-data/CodeGen scores 11.5/12.1; matching alone does not isolate data verification.
ScienceAgentBench (Chen et al., 2025d)	Task instructions, data, optional expert knowledge; self-contained Python and saved outputs.	Valid execution and task success; numerical/prediction checks, some visual model scoring; code rubric.	102 tasks, nine domain experts; three attempts; hidden labels/revised data where applicable.	Best-of-three is not single-run success; human code rubric can mix reference implementation with partial completion.
DSEval (Zhang et al., 2024)	Queries, reference code, files, and session state; code and updated runtime state.	Configurable validators check results, presentation, side effects, and execution; pass rate.	825 questions in 294 groups; default reference-state resets versus error propagation.	Ignoring presentation/integrity changes the success definition; error propagation is separately measured.
DSAEval (Sun et al., 2026)	Persistent multimodal notebook sessions; data, soft references, reasoning, code, and reports.	Two model judges weight reasoning/code/results by 0.3/0.3/0.4; cost and judge error.	Three doctoral raters calibrate on Gemini outputs for all 641 questions; visual-observation ablation.	Judge calibration covers one generator; code score is not hidden-test success.

Table 5: Detailed protocols for executable research workflows and scientific coding.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Curie (Kon et al., 2025)	Computational experiment questions, constraints, code, plans, execution records, and conclusions.	LLM design/setup/conclusion checks; human implementation–plan alignment.	46 reported tasks; shared GPT-4o; five runs; expert judge check described without sample/agreement detail.	Design pass 97.9%, conclusion correctness 36.1%; tasks generally do not require iterative hypothesis revision.
Machine-learning experimentation and research engineering				
MLAgentBench (Huang et al., 2024)	13 tasks, workspaces, baselines, file/code tools; final predictions or timed code.	Task-native metrics; success usually requires 10% relative improvement; time/token measures.	Eight runs; usually 50 actions/five hours, but GPT-4 has 30 actions.	Improvement means use of specified baseline; appendix CIFAR example exposes test feedback during tuning.
MLE-bench (Chan et al., 2025)	75 offline competitions; descriptions/data; prediction submissions, hidden-test scores, and logs.	Local task graders, valid submissions, historical medal thresholds, and pass@k.	24 hours/one A10 in main setup; sample-solution split checks; separate contamination/log audits.	o1-preview/AIDE 16.9% medal rate; pass@8 34.1%. Reconstructed tests limit exact historical equivalence.
RE-Bench (Wijk et al., 2025)	Seven optimization environments, weak starts, protected graders, expert/agent code and timed traces.	Task metrics normalized to start/reference solutions; best-score summaries and uncertainty.	61 experts, 71 eight-hour attempts; matched environments and budget-based comparisons.	32-hour human result is four independent attempts; selected noisy scores need re-evaluation for stable quality.
MLRC-Bench (Zhang et al., 2025b)	Seven research competitions; protected repositories; development-selected snapshots and hidden-test results.	Improvement relative to baseline/human solution; time and logical code lines; separate novelty study.	Usually eight attempts, 50 steps/five hours; task-specific test budgets; optional external ideas separated.	Scores may be negative. Novelty/effectiveness correlation -0.06 does not validate novelty judgments against novelty truth.
MLGym (Nathani et al., 2025)	13 tasks; code, weights, or policies; intermediate attempts and final submissions.	Task metrics; relative-performance profiles and area under those curves for best-of-four outputs.	Shared agent interface; four seeds; protected graders; repeated score feedback.	Best attempt and best submission are separate aggregates, not per-run best-solution retention rates.
FML-bench (Zou et al., 2026)	Eight research repositories; protected graders; iteration code and performance traces.	Task metrics; code-embedding dispersion; successful/executed and executed/allocated step rates.	Three rounds of 100 allocated steps; multiple agent/model combinations; failures recorded.	Code dispersion is not independently calibrated to hypothesis diversity; the original study's Table 2 ranks lower privacy scores as better, whereas its Appendix A defines the target as attack AUC near 0.5. These criteria differ if the tabulated scores are untransformed AUC.

Table 5 (continued): Detailed protocols for executable research workflows and scientific coding.

B.4 Academic Writing and Scholarly Communication Protocols

The records separate source-supported prose and aspect-level scoring, bibliographic validity and contextual citation support, and visual communication. Separate rows distinguish protocols within one work; calibration samples are not counted as additional works.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Source-supported writing and aspect-level evaluation				
CiteBench (Funkquist et al., 2023); citation text	Given cited papers and citing context; original citation sentences or paragraphs are references.	ROUGE, BERTScore, citation-intent/discourse KL; human readability and consistency.	Three annotators rate 150 outputs; 50 overlapping cases assess agreement.	Some reference targets require absent input facts. Authors document examples; affected prevalence is unquantified.
SciIG (Garg et al., 2026); introductions	Target title/abstract and cited metadata; generated and original introductions.	Reference-set precision/recall, keyword overlap, text metrics, model and human rubrics.	20 prompts, 80 outputs; three authors divide annotation without overlap.	3,900 tasks. Keyword factuality proxies align weakly with human ratings; source-set recovery is separate.
Task exploration (Sahinuç et al., 2024); citation paragraphs	Abstracts, reference-derived intent, and citation examples; generated paragraphs and atomic reference facts.	Text similarity, NLI, surface compliance; human atomic-fact coverage.	Two models; six prompts and six input settings; three annotators, 30 human-study examples.	Input manipulation is explicit. Automatic metrics only partly track atomic-fact coverage in this exploratory study.
GREP (Şahinuç et al., 2026a); related-work constraints	Main/cited abstracts and introductions, author preferences; generated sections and revisions.	Hard citation, support and positioning checks; soft length, emphasis and positioning-style scores.	50 balanced examples per criterion; 16 experts, three rounds each; separate synthetic and interactive checks.	Whole-section faithfulness requires every citation to pass. Parsing failures and semantic errors are not isolated in that rate.
OARelatedWork (Docekal et al., 2026); related-work generation	94,450 tasks with linked sources; validation/test require full texts; generated sections and statements.	Text similarity, citation F1; True/Wrong Citation/Unverifiable/False proportions; preference scores.	40 papers, 120 sections; 408 statements including adversarial cases; judge meta-evaluation.	Full-text input changes attribution errors. Unverifiable is not False; selected judge's wrong-citation F1 is .23 in the pre-extracted test.
STRoGeNS (Nishimura et al., 2024); paragraph organization	Target/cited titles and abstracts; paragraphs treated as citation clusters.	Cluster F1, ARI/ARI', paragraph-count error, missing-citation rate, ROUGE.	Four people rate 25 Llama2 outputs; Fleiss' $\kappa = .156$.	Human calibration concerns whole-section structure; evidence for paragraph-level measurement remains limited.
STRoGeNS (Nishimura et al., 2024); novelty statements	Generated related work; original and statement-deleted texts for validation.	GPT-4 detects statement presence; generation study samples 100 outputs.	190 original/constructed cases; reported detection accuracy 92%.	Measures presence of a novelty statement; authors leave its validity for future work.
SciXGen (Chen et al., 2021); object descriptions	Scientific object plus preceding text; author prose describing tables, figures, algorithms or theorems.	BLEU, METEOR, MoverScore, perplexity; 1–3 human ratings.	Five CS annotators; 50 samples per task; object-only, local and retrieved-context controls.	Description targets are prose spans, not captions. Context benefits vary by object type.

Table 6: Academic writing and scholarly communication protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
SciXGen (Chen et al., 2021); introduction paragraphs	Paper abstract and cited abstracts/full-text passages; original introduction prose.	Text metrics, human ratings; proportion of supplied references not mentioned.	Abstract-only versus cited-abstract/full-text inputs; CV subset.	The named 'Hallucination' percentage measures unmentioned references, supporting a coverage conclusion rather than a factual-error rate.
SciRM (Şahinuç et al., 2026b); aspect scoring	Query, explicit rubric and score examples; related-work/review-utility aspect labels.	Accuracy/F1 and repeated-run variability for rule-conditioned scores.	Human and synthetic labels; masked aspects; first-stage versus reflection training.	65,357 aspect-level instances, not independent papers. Four-task means weight tasks equally.
SciRM (Şahinuç et al., 2026b); unseen-task transfer	Novelty-assessment summary pairs and scientific revision examples.	Label accuracy/F1 on agreement and revision evaluation.	76 novelty pairs, 56 agreement labels; 6,184 revision aspect instances.	Novelty agreement accuracy near .74 is close to the majority baseline; the original study's Table 13 reports F1 without specifying the positive class or class-averaging convention.
Bibliographic validity and citation support				
Walters–Wilder (Walters and Wilder, 2023); generated references	84 generated reviews on 42 topics; 636 references checked against databases/publishers.	Manual existence, substantive metadata, formatting and link checks.	April 2023 GPT-3.5/GPT-4 outputs; suspected fabrications receive further source checks.	Fabrication: 55%/18%; errors among real works: 43%/24%. These have different denominators.
Chelli et al. (Chelli et al., 2024); review-search references	11 shoulder-disease systematic reviews, 109 included references; 362 model-generated entries.	Manual metadata checks, hallucination proportion, retrieval precision/recall.	Three 2023 models; two prompts per task, retaining the larger reference output.	Hallucination requires two of title/first author/year to be wrong. GPT-4 recall count is inconsistent (15 versus 16).
CiteAudit (Shi et al., 2026); metadata verification	Generated set: 6,086 entries; real-world set: 3,356; search-backed valid/invalid labels.	Error-positive confusion matrices, accuracy, precision, recall and F1.	At least two authors independently verify labels; unresolved disagreements excluded.	Reported real-set precision is .749 and recall is .951. Similar recall for one tool does not establish equivalent error distributions across sets.
CiteTracer (Li et al., 2026b); synthetic field verification	Perturbed BibTeX/PDF references; Real, Potential and Hallucinated codes and evidence links.	One-versus-rest F1; per-code true-/false-positive rates.	Triple validation; 2,450 BibTeX entries, 2,392 after PDF-rendering losses; five styles.	11 of 12 taxonomy classes tested. P3's matched-real-work definition conflicts with its pure-fabrication construction description.
CiteTracer (Li et al., 2026b); high-risk real references	957 preselected positive references from rejected conference papers; adjudication evidence.	Recall of Hallucinated labels; Potential reported separately.	Conference-chair decisions support the selected real-world cases.	929/957 labeled Hallucinated; flagging remaining cases as Potential is not the same classification endpoint.
Naser (Naser, 2026); reference audit	Non-retrieval outputs from ten models/four fields; 69,557 reported references.	Database matching, thresholded confirmation rates, repeated-output agreement.	CrossRef, OpenAlex and Semantic Scholar; automated search rechecks 75 cases per stratum.	Unconfirmed includes some real references. Authors treat the rate as an upper bound under this matching definition.

Table 6 (continued): Academic writing and scholarly communication protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Naser (Naser, 2026); string-only detection	27 bibliographic string features; confirmed/unconfirmed reference labels.	AUC, average precision and accuracy.	Five-fold stratified validation; leave-one-model-out evaluation.	GBM AUC .876; mean leave-one-model-out AUC .834. Labels inherit the audit's matching definition.
BibTeX errors (Rao and Callison-Burch, 2026); reconstruction and repair	Known title/first author; 931 papers, 2,793 entries; nine fields and multi-version metadata.	Six field states; field/entry correctness and repair regressions.	521-field human reannotation; native-search, one-stage and two-stage tool conditions.	The evaluation uses first-author matching; stricter author checks differ. Authors report loss of original data, preventing exact numerical reproduction.
HalluCitation Matters (Sakai et al., 2026); publication audit	17,842 ACL-family papers; 741,656 references screened; candidates manually checked.	Paper-level count with at least one confirmed hallucinated reference.	Source/PDF/identifier checks; stop checking a paper after one confirmed error.	295 papers (1.65%) is a lower bound on affected papers, not a complete citation-level error count.
GhostCite (Xu et al., 2026); reference generation	13 models, 40 CS areas, requested lists; 331,809 extracted references.	Validity proportion and JSON compliance reported separately.	Search/reasoning switches; manual checks of 400 valid and 400 invalid predictions.	The expected 375,440 entries is not the evaluated denominator; unparsed outputs are excluded from validity rates.
GhostCite (Xu et al., 2026); reference classification	50 valid and 50 invalid entries; models judge bibliographic identity.	Accuracy and class-conditional detection.	Single-entry prompts with search/reasoning enabled.	The column named 'Invalid Precision' divides by 50 actual invalid entries and is recall, not alarm precision.
GhostCite (Xu et al., 2026); archive audit	56,381 papers; 2,199,409 references; retrieval candidates and PDFs.	Confirmed invalid references and affected-paper proportions.	16 assistants; each candidate checked at least twice, with researcher review.	The original study's Section VI-A reports 739 invalid citations, versus 738 in its Table V. The authors state that the annual trend does not establish causality (their Section VI-D).
GhostCite (Xu et al., 2026); researcher survey	94 retained questionnaires about AI use and reference-checking habits.	Question-specific response proportions.	97 responses, three inconsistent responses excluded; opt-in recruitment.	Self-reports do not measure experimentally observed invalid-reference insertion rates.
Zhang–Abernethy (Zhang and Abernethy, 2024); quotation errors	250 natural statement–reference pairs; title, abstract, retrieved passage or PDF evidence.	Fully, partially or unsubstantiated accuracy; separate collapsed binary analysis.	Existing expert case labels; source-information and model controls.	Three-class accuracy 49.6–70.0%; more information does not uniformly help. Relevant subclaims define each source's responsibility.
BibAgent (Li et al., 2026c); open-source miscitation	Citing statement and source full text; 6,350 expert-checked synthetic errors.	Acc-pass@3 requires validity plus accepted explanation; abstention counts as error.	Same-backbone full-text controls; taxonomy annotation and over 3,000 grader decisions checked separately.	All main-set entries are errors. Pass@3 is not single-attempt deployment accuracy or false-alarm measurement.

Table 6 (continued): Academic writing and scholarly communication protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
BibAgent (Li et al., 2026c); inaccessible sources	Downstream citing contexts substitute for unavailable source full text; 3,810-error subset.	Acc-pass@3, abstention and conditional token savings.	Same-backbone search control; witness-count/consistency thresholds.	Different evidence and subset from Open. Savings use jointly non-abstaining cases; indirect witnesses may share misconceptions (author limitation).
Posters, slides and scientific diagrams				
Paper2Poster (Pang et al., 2025); poster quality	100 paper–author-poster pairs; generated single-page posters.	AltCLIP, perplexity and six GPT-4o rubric dimensions.	Author-poster and generation baselines; five repeated evaluations of one system's 100 outputs.	AltCLIP is not direct aesthetic judgment. Single-system repeatability does not establish ranking stability across systems.
Paper2Poster (Pang et al., 2025); PaperQuiz	100 questions per paper, split literal/interpretive; readers see only the poster.	Six VLM readers; raw QA and separately length-adjusted QA.	Full-paper/author-poster controls; one PhD participant, five papers, eight artifact conditions.	Length-adjusted score ranges to 200. Small human check limits the claimed general reader-proxy validation.
P2P/P2PEval (Sun et al., 2025a); poster content	121 paper–poster pairs; 1,738 author-poster-derived checklist items.	GPT-4o weighted checklist score, normalized to 0–100.	Three checklist annotators plus a fourth checker; visual items explicitly represented.	Content preservation and generic layout scores can change in opposite directions under reasoning-mode changes.
P2P/P2PEval (Sun et al., 2025a); generic poster quality	Ten generic dimensions and human overall scores; generated posters.	Ten 0–5 dimensions; XGBoost calibration; separate pairwise preference.	1,701 human scores; ten-fold validation reports $R^2 = .92$.	Calibration predicts human totals; the original study's Section 3.2.1 does not specify whether folds are grouped by paper, poster, or individual rating. Strict wins and wins-or-ties differ.
SlideTailor (Zeng et al., 2025); preference-conditioned slides	Target paper, example paper–slides pair and PPTX template; editable ten-slide output.	Topic IoU, ordering, structure, preference and generic quality; six normalized scores.	50 main cases; four people supply 60 ratings on 30 cases; cross-judge checks.	40 wins/9 losses/11 ties: 81.63% excludes ties. Potential 100,000 combinations are not generated test cases.
AutoFigure (Zhu et al., 2026b); scientific illustrations	Long scientific text and reference illustration; 300 test cases within 3,300 pairs.	Eight VLM submetrics; automatic blind comparisons; human ratings/rankings.	Ten first authors evaluate 21 papers; separate 95-output human–judge calibration.	The original study's Table 2 reports automatic evaluation; human evaluation is reported separately. Authors document residual text errors after correction; human adoption denominator is unclear.
PaperBanana (Zhu et al., 2026a); method diagrams	Method text/caption and reference library; 292 reference and 292 test cases.	Four reference-comparison dimensions; hierarchical overall win/tie/loss score.	50 judge-calibration cases; separate 50-case three-person blind preference study.	Reference baseline 50 is defined by scoring. Overall 60.2 coexists with faithfulness 45.8; authors acknowledge this shortfall.

Table 6 (continued): Academic writing and scholarly communication protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
PaperBanana (Zhu et al., 2026a); statistical charts	Underlying table data; Matplotlib-generated charts; ChartMimic-derived cases.	Reported chart-quality comparison under the chart protocol.	240 reference and 240 test charts across seven types; baseline comparison.	Separate code-rendered task, not the method-diagram test; reported overall improvement is 4.1 points.
SciFig (Huang et al., 2026); editable method diagrams	Method/caption to draw.io XML, including embedded PNG components; 435 tests.	50-question VQA, visual/OCR similarity, 20-item rubric and eight design heuristics.	1,784 rubric-development cases; 60 people, 10,000 blind comparisons; human alignment checks.	Supports component-level editing through XML with embedded PNG assets; human comparisons assess preferences for generated diagrams.
SridBench (Chang et al., 2025); scientific illustrations	1,120 image–caption–section triples across 13 subject directions.	Six GPT-4o 1–5 dimensions relative to original diagrams.	100 cases compare two model judges with human ratings; two generators fully evaluated.	Authors report textual omissions and scientific errors. Low means do not define an absolute zero-capability threshold.

Table 6 (continued): Academic writing and scholarly communication protocols.

B.5 Automatic Peer Review Protocols

The records distinguish review-content evaluators, manuscript-grounded criticism, multimodal checking, and feedback use or aggregation. Natural weaknesses, injected errors and decision labels supply different references; a paper may therefore contribute several task-specific rows.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Review content, evaluators and calibration				
ReviewEval (Garg et al., 2025); review dimensions	Manuscript, official reviews and venue rules; generated reviews and automatic rebuttals.	Semantic/topic alignment; question-level support; depth, constructiveness and guideline scores.	120-paper corpus; five experts compare outputs on a 30-paper subset across four preference dimensions.	Preference calibration does not independently label factual support decisions. ReviewAgent is a separate evaluated generator.
GEM/GRE-bench (Xu et al., 2025a); shared information	Candidate and peer-reference reviews; optional abstract/author-strengths context.	Probability-model estimate of shared information; GEM-S conditions out supplied background.	Teacher correlations on about 180 proposal reviews; separate degradation/manipulation tests on 300 ICLR papers.	Requires peer references, not unique gold. GEM-S teacher correlation .479 is below LMExaminer's .537; robustness tests measure another property.
Beyond Rating (Li et al., 2026a); multidimensional review alignment	1,000 papers, 3,666 high-confidence/low-disagreement reviews; extracted atomic points.	Summary coverage, point precision/max-recall, focus KL, QuestionScore, Binoculars and rating MAE.	Alternative extraction-model check; shared prompts for generic models.	Binoculars lacks independent depth calibration; QuestionScore checks retained text after removing some sections.
MindA (Shin et al., 2025); focus alignment	Review/meta-review points coded by manuscript target and evaluation aspect.	Group-distribution KL; per-paper target–aspect F1; text-similarity controls.	327 points from 68 papers compared with author-discussion labels; near-submission versions for accepted-paper supplement.	Focus means category frequency, not attention weights or critique truth. Abstract/main-text sample counts differ.
Liang et al. (Liang et al., 2023); feedback overlap	Nature/ICLR papers and reviews; GPT-4 comments from the first 6,500 input tokens.	Extracted-point overlap, hit rate and matching coefficients.	Human extraction/matching validation; shuffled-paper specificity control; matched point-count comparison.	Reported overlap measures shared feedback concerns. Rejected-paper overlap and total review counts have version-internal conflicts.
Liang et al. (Liang et al., 2023); perceived usefulness	Uploaded manuscripts, model feedback and questionnaires from 308 researchers.	Perceived helpfulness and comparisons with prior human feedback.	Prospective opt-in users from AI/computational biology; user survey separate from corpus matching.	57.4% rate feedback helpful/very helpful. The 82.4% comparison is with at least some human feedback, not the average expert.

Table 7: Automatic peer-review evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Zhou et al. (Zhou et al., 2024); rating inference	PeerRead paper text or human review text; aspect/overall reference scores.	Score error and correlation.	Paper-input versus review-input conditions; zero-/few-shot prompts.	Rating recovery from reviews and assessment from manuscripts yield different correlations.
Zhou et al. (Zhou et al., 2024); review generation	ASAP manuscripts to aspect/sentiment-annotated reviews; human reference reviews.	ROUGE, BERTScore, BLANC; model and human relevance/informativeness ratings.	300-paper GPT-3.5 sample; 50-paper GPT-4 human analysis.	BERTScore has near-zero reported correlation with human relevance/informativeness in this sample.
Zhou et al. (Zhou et al., 2024); RR-MCQ	196 multiple-answer questions from 14 papers' review–rebuttal discussions.	Per-option accuracy and whole-question exact correctness.	Shuffled options; section-extraction variants; disputed source discussions excluded.	GPT-4 pipeline: .710 per option, .276 per question. Reported F1 conflicts with the listed precision and recall.
Manuscript-grounded criticism and error finding				
CLAIMCHECK (Ou et al., 2025); claim association	Manuscript plus human weakness; locate up to three directly criticized claim spans.	Edit/exact-span F1.	Expert links; pilot agreement separate from mainly single-annotator main set.	41 papers, 168 weaknesses; 120 weaknesses linked to 154 target claims. Pilot human scores are not same-test upper bounds.
CLAIMCHECK (Ou et al., 2025); weakness labeling/editing	Known weakness and target claims; classify and rewrite the criticism.	Krippendorff's alpha for labels; rewriting preference and meaning change.	20 rewrites per model; two judges each, randomized order.	Supplying the weakness distinguishes this task from independent discovery; preferred rewrites may change meaning.
CLAIMCHECK (Ou et al., 2025); claim verification	Manuscript and claim, without supplied weakness; generate/check criticism.	Label agreement; Same, Similar or Different reference-weakness matching.	Compared with weakness-given condition using the same task family.	Low reference agreement is reported; a different criticism is not automatically an incorrect one.
Peerispect (Ghorbanpour et al., 2026); CMC	500 extracted manuscript claims from 50 ICLR papers; all supported by construction.	Supported-label recovery accuracy; evidence retrieval recall.	Retriever/reranker and 3B/7B verifier comparisons; top-three passages.	Single-positive-class accuracy does not test support-versus-nonsupport discrimination; retrieval is separately measured.
Peerispect (Ghorbanpour et al., 2026); RRC	150 real review claims from 25 papers; human labels informed by manuscript and rebuttal.	Four-class verification accuracy.	Same modular comparisons; no unique oracle paragraph, hence no RRC retrieval recall.	Best reported four-class accuracy is .287 on this real-review set.
FLAWS (Xi et al., 2025); injected-error localization	713 paper–error pairs; PDF, injected claim-invalidating edits and associated reference spans.	Top-k localization by span similarity or semantic match; bootstrap intervals.	58-error human audit; 253 agreed detection judgments calibrate scoring; insertion-source regression controls.	57/58 sampled errors valid; scorer precision .91/recall .98 on agreed cases. k is ranked candidate depth, not reruns.
SPOT (Son et al., 2025); acknowledged-error discovery	83 historical papers with 91 author-acknowledged, self-contained errors; full text/images.	Semantic reference matching; precision, recall and pass@K.	Eight runs per paper/model; 48-paper text/image comparison; domain-expert case audits.	Reference-outside true error documented. Unmatched findings count as false positives, so matching precision is not critique truth rate.

Table 7 (continued): Automatic peer-review evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
LimitGen (Xu et al., 2025b); synthetic limitations	500 NLP papers, 1,000 manually checked perturbation cases; given target aspect.	Top-three limitation identification; match-gated specificity/relevance scores.	Human correction of 112 cases; sampled human checks; with/without retrieval.	Perturbation cases are not 1,000 independent papers; authors limit the scope of cross-domain extension.
LimitGen (Xu et al., 2025b); natural limitations	1,000 ICLR papers and extracted review weaknesses; aspect-conditioned output.	Reference precision/recall/Jaccard and match-gated quality; independent human dimensions.	Human faithfulness, soundness and importance ratings initially hide references; retrieval comparison.	Unmatched plausible limitations receive automatic zero. Human quality evidence is separate from reference coverage.
ReviewScore (Ryu et al., 2026); claims and questions	Review points plus manuscript text/tables; expert factuality and answerability labels.	Misinformed-detection F1 and five-point weighted kappa.	Three-person annotation; rebuttal-input comparison.	Main set has 573 points from 40 papers; the 657-point total includes pilot annotations. Questions check whether the paper already answers them.
ReviewScore (Ryu et al., 2026); argument premises	Arguments reconstructed into premises/conclusion, formal logic and natural language.	SAT validity; premise factuality; conjunction or weighted aggregation.	Independent human reconstruction faithfulness; low-faithfulness reconstructions gated out.	Authors report aggregation can propagate errors or reduce agreement; logical validity and premise truth are separate checks.
Counterfactual review (Dycke and Gurevych, 2026); reasoning-sensitive response	133 papers, 391 critical and 540 neutral edits; generated reviews and ratings.	Average treatment effects on review aspects, positive assertions and scores; mixed-effects tests.	Neutral rewriting control; oracle positive control; 92 edits from ten papers human-checked.	No significant extra critical-edit response in tested statistics. Full-pipeline response does not isolate reasoning from visibility/localization.
Multimodal manuscript evidence				
PRISMM-Bench (Selch et al., 2026); identification	384 verified text/figure/table inconsistencies from 353 manuscripts; four-choice questions.	Accuracy; visual reliance relative to no-context performance.	Focused, page and document inputs; natural-language versus JSON options; matched human subset.	Focused evidence is already localized. Structured options reduce, but do not eliminate, answer-wording shortcuts.
PRISMM-Bench (Selch et al., 2026); remedy selection	384 inconsistencies; Target–Action remedy options and manuscript evidence.	Four-choice accuracy under three context levels.	Target–Action JSON options converted from manually refined identification choices.	Tests choosing a repair, not executing one; authors explicitly define this endpoint.
PRISMM-Bench (Selch et al., 2026); visual matching	192 visual–visual inconsistency pairs; focused crops and candidate matches.	Four-choice matching accuracy.	Focused context only; underlying evidence manually verified.	The total of 960 task instances includes repeated source inconsistencies across the three protocols.
MMReview (Gao et al., 2025); summary (S)	240 manuscripts; generated summaries versus human-review summaries.	BARTScore and LLM semantic/content comparison.	Text, text+figures and page-image conditions.	Task asks for a summary in the model's own words, excluding subjective assessment.

Table 7 (continued): Automatic peer-review evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
MMReview (Gao et al., 2025); strengths (SE)	238 manuscripts; generated acceptance-oriented strengths versus human points.	BARTScore and LLM reference comparison.	Shared modality controls; manually verified reference extraction.	Quality, clarity, significance and originality guide the output.
MMReview (Gao et al., 2025); weaknesses (WE)	240 manuscripts; generated rejection-oriented criticisms versus human points.	BARTScore and LLM reference comparison.	Same modality comparisons as summary/strength tasks.	Scores concern coverage and semantic similarity to selected human criticisms.
MMReview (Gao et al., 2025); soundness (SS)	32 manuscripts with human soundness scores; model assesses technical/evidential support.	MAE on integer 1–4 scores.	Modality and model comparisons.	This task's 32 cases are a subset, not the full 240-paper corpus.
MMReview (Gao et al., 2025); presentation (PS)	32 manuscripts with human presentation ratings; clarity and organization.	MAE on integer 1–4 scores.	Modality and model comparisons.	Presentation and technical soundness have separate reference ratings.
MMReview (Gao et al., 2025); conditional decision (CD)	176 cases with human-written reviews supplied; infer overall paper score.	MAE on 1–10 ratings.	Compared with paper-input decision tasks, retaining the input difference.	Extra human reviews supply evaluative information absent from DD/CoD.
MMReview (Gao et al., 2025); direct decision (DD)	176 manuscripts; directly produce overall score.	MAE on 1–10 ratings.	Direct-input baseline for structured CoD.	Direct paper assessment differs from review-conditioned score recovery.
MMReview (Gao et al., 2025); structured decision (CoD)	176 manuscripts; summary, strengths/weaknesses, subscores, then overall score.	Overall-score MAE.	Same task family as DD; structured intermediate reasoning.	Improvement over DD varies by model; Tiny/Middle group means are worse in the reported table.
MMReview (Gao et al., 2025); meta decision (MD)	240 cases with multiple human reviews; binary accept/reject output.	Decision accuracy against recorded outcomes.	Guided review synthesis rather than only averaging numerical scores.	Paper-alone decisions and human-review aggregation use different evidence.
MMReview (Gao et al., 2025); pairwise rank (PR)	96 paper pairs from different acceptance tiers; choose relative preference.	Accuracy against recorded tier ordering.	Each pair repeated with reversed input order.	Preference target is conference-tier alignment.
MMReview (Gao et al., 2025); fake strengths (FS)	240 cases containing polarity-reversed human-review claims; manuscript evidence.	MAE as specified in the original study's Table 2 for numerical evaluations.	GPT-4o antonymic rewrites with manual semantic checks.	Constructed false-positive assessments, separate from open-ended strength generation.
MMReview (Gao et al., 2025); fake weaknesses (FW)	238 cases with polarity-reversed human-review claims; manuscript evidence.	MAE as specified in the original study's Table 2.	Same construction and manual verification family as FS.	Constructed criticism assessment, not an unrestricted error-discovery task.
MMReview (Gao et al., 2025); prompt injection (PI)	176 manuscripts with hidden instructions; otherwise the CoD setup.	Score deviation, increase/decrease frequency and conditional change magnitude.	Clean-input control; PDF-as-image excluded from this text-injection task.	Change among increasing cases is not average change over all cases.

Table 7 (continued): Automatic peer-review evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Usefulness, intervention and opinion aggregation				
RevUtil (Sadallah et al., 2025); author-oriented review utility	Review point only, without manuscript; 1,430 human-labeled points and synthetic training labels.	Four 1–5 aspects; weighted kappa, correlation; claim-detection F1.	Three annotators per point; full human set and majority-agreement subsets reported separately.	Grounding is specificity of reference to the paper; verifiability is provision of reasons. Neither checks those reasons against the manuscript.
Thakkar et al. (Thakkar et al., 2025); randomized feedback	ICLR review versions, random assignment, delivered feedback and subsequent discussion.	Edits, length, uptake, discussion and acceptance statistics.	22,467 selected/22,364 control; intention-to-treat separate from received-feedback analysis.	18,946 actually received feedback after gating. Assignment and receipt have different denominators.
Thakkar et al. (Thakkar et al., 2025); revised-review preference	100 before/after review pairs with 3–4 feedback items and uptake above .60.	Blind researcher preference for revised versus initial review.	Two AI researchers; selected high-uptake revisions.	89% preference applies to this selected subset, not every assigned or delivered review.
ORSUM (Zeng et al., 2024); meta-review generation	Title, abstract, review set and given decision; generated consensus/dispute synthesis.	Text/factuality metrics; model rubrics; human binary informativeness, soundness, consistency and faithfulness.	Three annotators; 50 difficult borderline cases; three automatic runs.	Reviews capped at 300 tokens; no rebuttal. Decision is an input; human/LexRank perfect faithfulness is assumed, not measured.

Table 7 (continued): Automatic peer-review evaluation protocols.

B.6 End-to-End Research Protocols

System demonstrations and review-feedback studies are followed by multi-task infrastructure, connected research subprotocols, and rediscovery. The suite, stage and full-run rows preserve their different denominators.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
System demonstrations and feedback				
AI Scientist (Lu et al., 2024); connected research	Baseline code, plot scripts and LaTeX templates; ideas, code changes, experiments, plots and papers.	Task-native losses/KL; funnel counts; automated review scores and costs.	Three lightweight ML domains; multiple backbones; author case inspection.	Experiment completion and compiled-paper counts differ. Reported threshold crossings are automated-review outcomes.
AI Scientist (Lu et al., 2024); internal reviewer calibration	500 ICLR 2022 papers and decisions; text-only model reviews.	Decision accuracy, balanced accuracy, F1/AUC and rating correlation.	Reflection, ensemble and threshold variants; published NeurIPS human statistics as a reference.	Cross-conference human comparison and model-to-mean versus person-to-person correlations do not establish same-condition human-level performance.
DeepScientist (Weng et al., 2025b); sustained research	Reproduced baselines, findings memory, implementations and task results across three AI tasks.	Accuracy, throughput, AUROC/latency; idea-to-validation-to-progress counts.	Human result verification; random-idea filtering comparison; failed-implementation audit.	About 5,000 ideas, 1,100 validations, 21 progress findings; 20,000+ GPU hours. Historical publication spans are not measured human research time.
DeepScientist (Weng et al., 2025b); generated papers	Five system papers and external generated papers; automatic and expert reviews.	DeepReviewer ratings; three-person simulated review panel and agreement.	Three independent human ratings per paper; no rebuttal.	Automatic mean rating 5.90 and human mean 5.00 are different channels. Simulated acceptance is not conference acceptance.
CycleResearcher (Weng et al., 2025a); simulated research feedback	Bibliography/abstracts to long research papers with simulated experiments; review-guided refinement.	Automated paper rubrics and simulated acceptance; selected expert review.	Independent reward-model check; three NLP experts compare selected outputs anonymously.	Main protocol explicitly simulates results. The original study's Appendix I separately demonstrates one executed study; it is not the main evaluation.
CycleReviewer (Weng et al., 2025a); review calibration	Review-5k paper/review data; generated reviews and predicted decisions.	Proxy MAE/MSE against other reviewers' means; decision accuracy/macro-F1.	Human-reviewer proxy baseline under n versus n-1 aggregation.	Agreement with aggregate ratings is the proxy target. Reported percentage reductions conflict with rounded table values.

Table 8: End-to-end scientific research system evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
Suites, research stages and task-native outcomes				
AstaBench (Bragg et al., 2026); multi-task suite	11 adapted benchmarks, 2,400+ problems; literature, code, analysis and discovery tasks.	Task-specific program/model scorers; macro aggregates and token-cost metadata.	Standard tools/environments; custom-agent tooling disclosed by category; frozen price table.	57 agent configurations are not a complete 57-by-11 matrix. Suite macro score is not a connected research-run success rate.
AstaBench E2E-Bench (Bragg et al., 2026); research steps	Expert-revised detailed research questions; reports, code and generated artifacts.	Rubric step passes require no conflicting evidence and at least one supporting source; all-step completion separately.	40 test/10 validation tasks; Faker negative control; human grader check of 50 development-set rubric items.	Human check: 92% grading correctness. The original study's Table 21 separately reports 16% report-apparent passes contradicted by code/artifacts. Faker's whole-task completion is zero.
AstaBench E2E-Hard (Bragg et al., 2026); harder research tasks	HypER-generated/refined questions without the same simplification; three artifact channels.	Mean step score and all-required-step completion; per-task cost.	Separate 40 test/10 validation tasks; same cross-artifact rubric logic and Faker control.	Panda step score 68.2 and whole-task completion .03 have different units; not the full-suite result.
MLR-Bench (Chen et al., 2025b); ideas and proposals	201 workshop-derived research tasks; ideas and proposals, with shared sampled upstream input in stage tests.	Phase-specific MLR-Judge rubrics; two model scores averaged.	Six generation models; retrieval between idea and proposal phases.	201 is the early-stage task count, not 201 completed end-to-end studies.
MLR-Bench (Chen et al., 2025b); experiments and writing	Ten selected implementable tasks; proposals, code, logs, figures, reports and papers.	Experiment soundness/completeness and paper rubrics.	Common stage inputs; coding/writing model comparisons; four RTX 3090 environment.	Authors find unverified synthetic/placeholder-data results in eight of ten Claude Code tasks; limited to that configuration.
MLR-Bench (Chen et al., 2025b); end-to-end and judge checks	Ten-task connected runs; final papers plus execution evidence; expert ratings.	Two-judge overall scores; human–model versus human–human absolute differences.	Ten experienced experts, two ratings per paper; experts mainly read papers.	Non-significant discrepancy-distribution tests do not establish equivalence. Human/model evidence permissions differ.
ResearchGym (Garikaparthi et al., 2026); research execution	Five repositories, 39 subtasks; source method removed, baselines/data/graders retained.	Task-native scores, agent/SOTA normalization, completion and improvement rates; means/best@3.	Three initial runs/task, 12 h, A100 80GB, $10 API; source-method reinsertion check; post-run integrity audit.	Main agent improves visible baseline in 1/15 runs; 26.5% mean subtask completion. Best-run extensions have additional budget.
AIRS-Bench (Lupidi et al., 2026); solution construction	20 problem/dataset/metric tasks, no initial solution code; prediction CSV on hidden test labels.	Native scores; valid submissions; normalized score and Elo; failures score zero.	At least ten runs per configuration and task; 24 h and one H200; human checks of task configuration.	Four tasks have at least one result above literature SOTA. Valid-submission rates conflict in text/figure; training exposure is not ruled out.

Table 8 (continued): End-to-end scientific research system evaluation protocols.

Work and task	Input/output evidence	Evaluator/metric	Calibration/controls	Findings/conditions
InnovatorBench (Wu et al., 2025); long-horizon LLM research	20 tasks/six categories; workspaces with core innovation removed; data, models or code submitted.	External task-specific metrics; normalized final/best scores, weighted aggregates and cost.	13 task builders reproduce references; common ReAct; at most three intermediate evaluations; hints disabled in main runs.	Reference solution is near 80, not 100. Best means best submission within a run. Per-task compute/time and network permissions vary.
Scientific rediscovery and diagnosis				
FIRE-Bench (Wang et al., 2026b); core rediscovery	High-level question/data/conditions; original implementation and findings hidden; generated conclusions versus reference findings.	Atomic-claim entailment precision/recall/F1, averaged over runs.	30 core tasks, three runs/system; human claim-extraction check; inspected trajectories for source access.	Best mean F1 46.7 is a claim score, not task success percentage. Added human calibration covers extraction, not the final entailment scorer.
FIRE-Bench (Wang et al., 2026b); cross-domain extension	Ten additional tasks from the 100-paper source pool; generated findings and reference claims.	Same claim-level scoring, separate cross-domain results.	Three systems; Claude Code excluded for budget reasons; repeated runs.	Other 60 source papers are parsed/released but not run in this study. All three reported means are below 25 F1.
FIRE-Bench (Wang et al., 2026b); failure attribution	Execution trajectories, code, conclusions and original paper; unmatched/incorrect findings.	Four stages, 16 error classes; false-positive categories including alternative conclusions.	LLM preliminary annotation followed by author item-wise review.	Authors identify planning/conclusion failures and occasional plausible alternative conclusions; reference-outside conclusions can be penalized.

Table 8 (continued): End-to-end scientific research system evaluation protocols.