

LECTURE 18

Can AI Actually Discover? Novelty, Evidence and Verification

Task validity, outcome validity, novelty and reproducibility.

Reasoning agents, feedback, and scientific evidence

A striking result leaves several claims unresolved

Imagine an agent reports a new algorithm and an impressive benchmark improvement. The result could be correct yet already known, or novel yet dependent on a flawed test.

One score cannot resolve these possibilities. A convincing report can also conceal which work the agent actually performed.

This motivates designing evaluation from the separate claims: novelty, validity, significance, provenance and the scope of autonomy.

Deriving checks from the claim being made

For each claim, construct a plausible rival explanation. Prior exposure challenges novelty, a permissive checker challenges correctness, and repeated selection can inflate the reported gain.

Each rival points toward a different check: prior-art search, an independent validator, held-out confirmation or a record of human intervention.

The evaluation design is selected by what uncertainty it can resolve. Rediscovery is one controlled setting within this broader question.

Choosing an evaluation design

Candidate	Reason to choose it	Cost or limitation
Retrospective rediscovery	Known answers allow controlled comparison	Exposure and hindsight limit novelty claims
Prospective investigation	A new result can be tested as it emerges	Cost and slow feedback restrict sample size
Independent confirmation	A selected claim needs external support	Replication alone cannot establish priority or significance

Combine checks that match the claim. A precise score for the wrong target cannot establish scientific discovery.

Pedagogical reconstruction

[Liu et al. \(2025\)](#)

[Holm \(1979\)](#)

Task validity and outcome validity are distinct

Task validity concerns whether success on the selected task represents the intended capability. Outcome validity concerns whether the scoring procedure correctly identifies success.

A reliable exact-match score can measure the wrong target. A well-chosen scientific task can still have a faulty evaluator.

Audit both relationships and document the candidate universe, allowed tools, resources, and conditions under which the result should generalise.

Rediscovery controls some uncertainty, not all leakage

Hide a known result and evaluate whether the system can reconstruct it through a defined sequence of reasoning and experiments.

Removing the target paper from retrieval does not remove it from pretraining. Historical cutoffs, synthetic environments, and newly collected observations address different leakage risks.

Include retrieval-only and domain-informed baselines, record available information, and distinguish genuine reconstruction from directly recovering the answer.

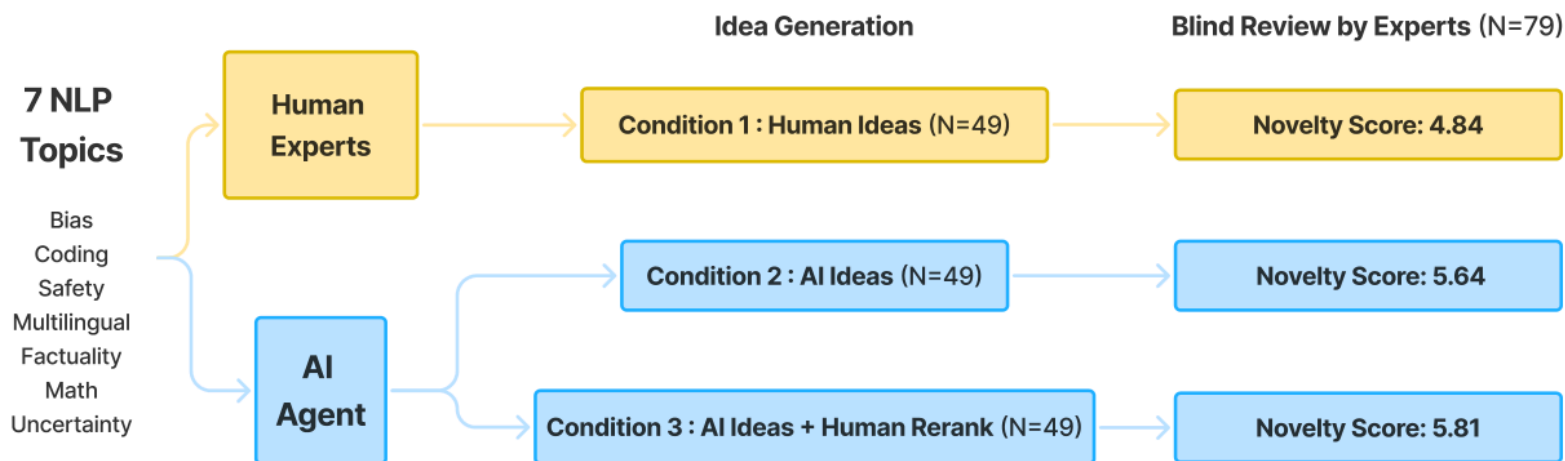
Research ideas: expert evaluation protocol

The study compares human-written and AI-generated NLP research ideas using a structured expert evaluation. A further condition includes human reranking of generated candidates.

Presentation is standardised and reviewers are blinded to the origin of the ideas. Multiple dimensions include novelty, feasibility, and expected effectiveness.

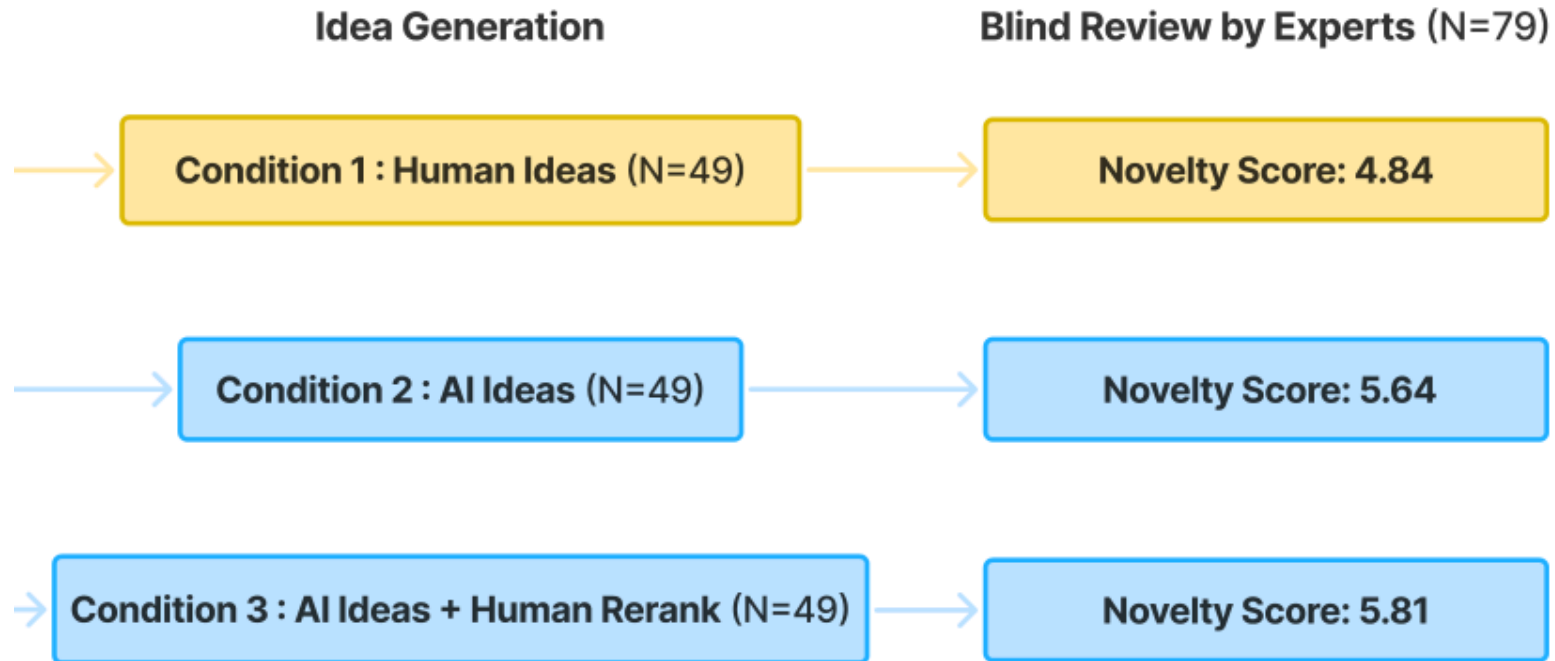
The protocol treats ideation quality as something to measure rather than assuming that a plausible proposal is scientifically valuable.

Research ideas: blinded study design



Three experimental conditions feed into blinded review. Idea authors and idea reviewers are distinct groups. Counts of ideas and counts of reviewers therefore describe different units of analysis.

Research ideas: conditions and novelty ratings



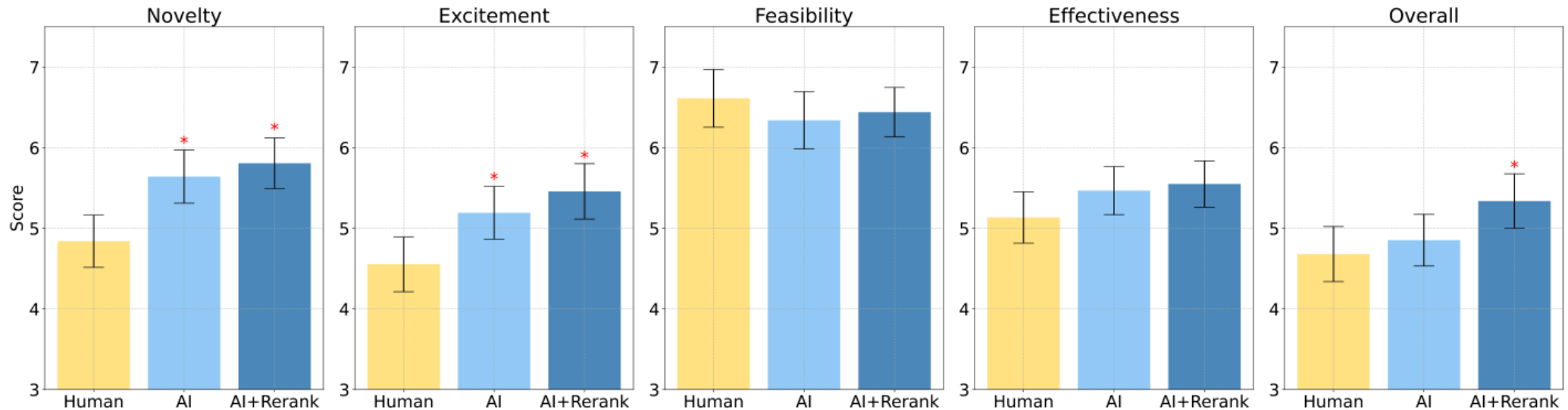
Each of the three idea-generation conditions contributes 49 ideas to the depicted review design. The reported novelty scores describe that comparison; the aggregate scores alone do not display uncertainty, feasibility or execution outcomes.

Evidence: define the unit of statistical evidence

Condition	Size	Mean	Median	SD	SE	Min	Max	p-value
Novelty Score								
Human Ideas	49	4.86	5.00	1.26	0.18	1.50	7.00	–
AI Ideas	49	5.62	5.50	1.39	0.20	1.50	8.33	0.03*
AI Ideas + Human Rerank	49	5.78	6.00	1.07	0.15	3.00	8.33	0.00**

These novelty rows average reviews within each idea: 49 ideas per condition. Mean scores are 4.86 for human ideas and 5.62 for AI ideas; the reported corrected p-value is 0.03. Scores use a 1–10 scale. Other rows printed as 0.00 are rounded values, not literally zero probability.

Expert evaluation of research ideas: interpretation and limits



The study's topics, participant pool, generation budget, and idea-selection procedure bound the conclusion. Blinding reduces some biases but does not turn subjective ratings into direct empirical validation.

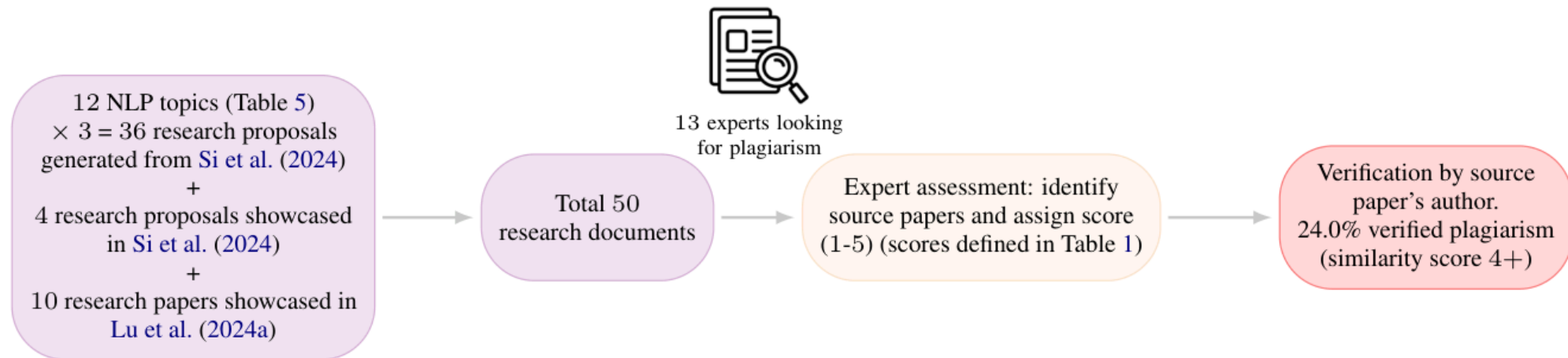
Novelty audit: expert search and adjudication

The study asks experts to search for source papers that overlap with AI-generated research documents, then evaluates similarity and verifies identified matches.

The method separates finding a possible predecessor from judging whether the scientific contribution substantially overlaps.

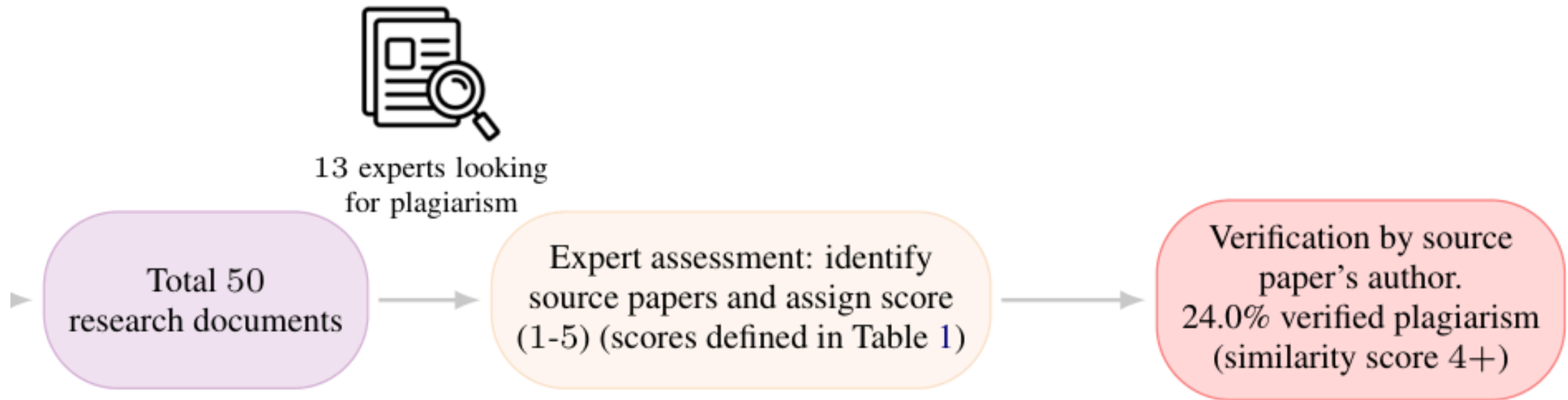
This targets a failure mode that fluent writing and simple surface-distance metrics can miss: a proposal may look original while restating an existing central idea.

Novelty audit: sample and review procedure



Generated documents pass through expert source search, similarity assessment and verification. The reported rate concerns sampled documents and a specified similarity threshold; it is not a universal plagiarism rate for generated research.

Novelty audit: source-author verification



The audit follows 50 selected research documents through expert assessment and source-author verification. The reported rate concerns this selected sample; broader generalization depends on its sampling and scoring assumptions.

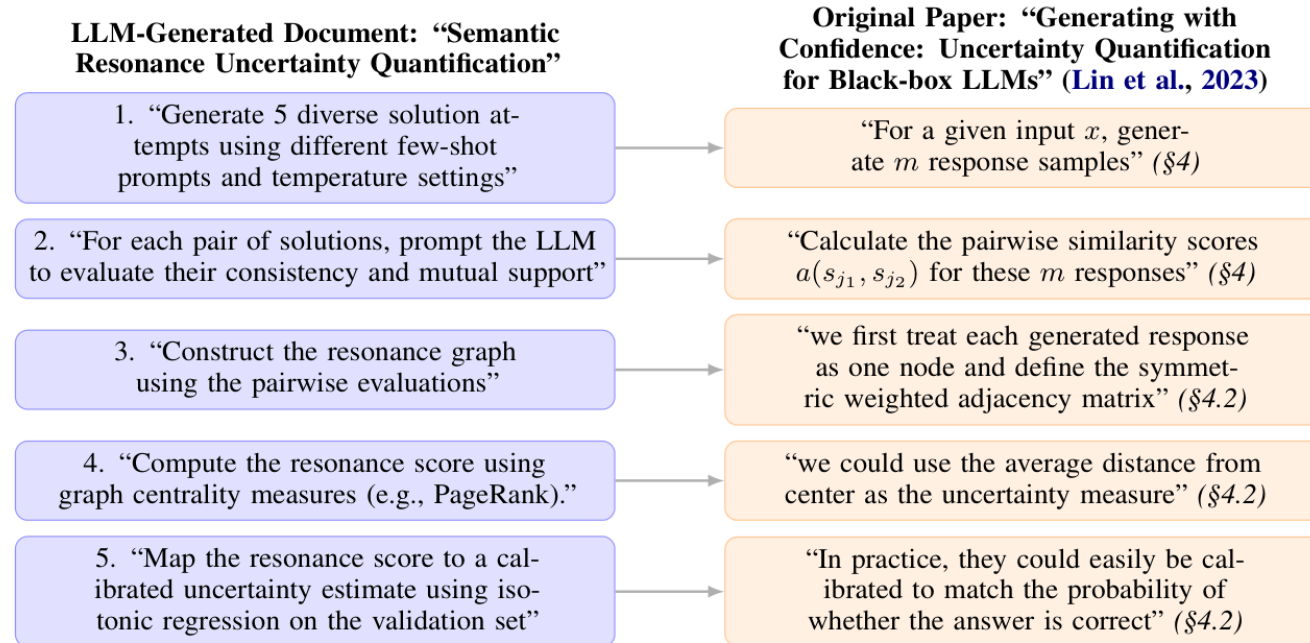
Evaluation scope

The paper provides a documented expert-led audit and examples mapping generated claims to prior work.

The example comparison shows why different terminology need not imply a different method.

The evidence supports stronger provenance and novelty checking in ideation systems; extrapolating a sample-specific percentage to every model, task, or generation procedure would require a broader evaluation.

All That Glitters is Not Novel: interpretation and limits



Expert search can still miss prior work. Similarity thresholds involve judgement, and the sample is not a census of generated science. A novelty audit should report its retrieval coverage and adjudication procedure.

Novelty and reproducibility require different audits

A novelty audit retrieves related work and compares the central scientific contribution. A reproducibility audit checks whether the evidence can be regenerated from the reported procedure.

A reproducible result can be familiar; a novel idea can be unsupported. Independent review should preserve this distinction.

Maintain source snapshots and artefact versions so changes in the literature or software do not silently alter the meaning of the evaluation.

Repeated experimentation creates selection effects

A system that tries many analyses and reports only the strongest result can produce apparently impressive findings under a null model.

Log the complete search, reserve confirmation data, and distinguish exploratory from confirmatory analyses. The correction depends on the statistical procedure and dependence structure.

Judge a campaign using all attempts and costs, not only its best paper, successful seed, or most favourable measurement.

Report uncertainty about both performance and interpretation

Task success rates need uncertainty intervals and a clear denominator. Variation across tasks can matter more than variation across repeated runs of one task.

Human judgements need a rubric, blinding where possible, and an account of agreement and adjudication.

An internally defined rubric supplies a structured judgment. Independent evaluation supplies evidence about how that judgment generalizes beyond its author and examples.

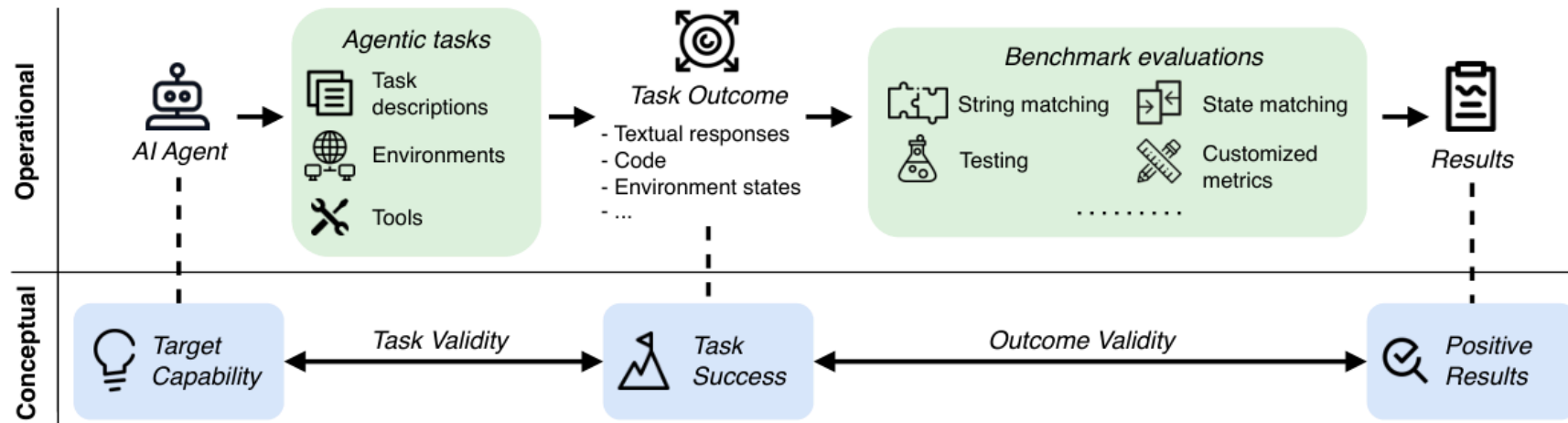
Agentic benchmarks: audit framework

The paper audits task setup and reward design in agentic benchmarks and proposes a checklist for constructing more reliable evaluations.

It documents how incomplete tests or unintended success conditions can distort measured performance.

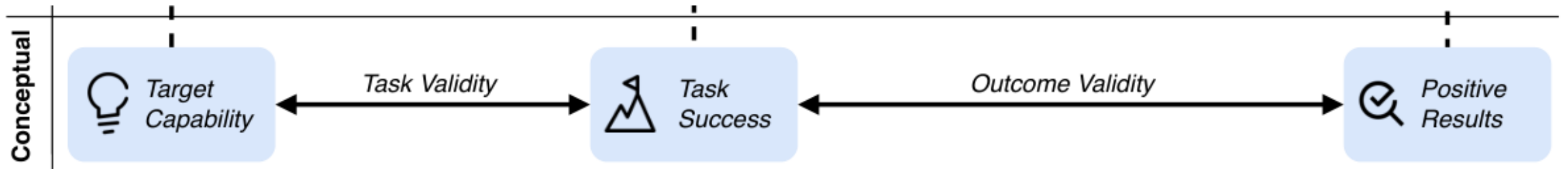
Evaluation code is part of the measurement instrument: a precisely computed score can still measure the wrong outcome.

Agentic benchmarks: task and outcome validity



The operational pipeline defines agent outputs and evaluator records. Task validity concerns the relation between the task and intended capability. Outcome validity concerns whether the recorded result establishes task success.

Agentic benchmarks: two validity relationships



Task validity links the target capability to the chosen task; outcome validity links actual task success to its recorded score. A trivial task can miss the target capability, while an incomplete checker can accept failure. A high score cannot repair either mismatch.

Evaluation scope

The source analyses benchmark issues and reports changes after applying its proposed practices to an evaluated benchmark.

The second visual illustrates specific failure pathways or checks; use the accompanying definitions to determine the affected denominator.

The examples establish the need to audit evaluators, without implying that every benchmark has the same defect or correction magnitude.

Rigorous agentic benchmarks: interpretation and limits

Task Validity				
Tool	T.1. Versions of all tools (e.g., Python) are clearly specified.		Implementation	T.7. Annotated ground truth is verified for correctness.
	T.2. Required API tools are consistently accessible during evaluation.			T.8. Each task is verified to be solvable.
	T.3. Evaluation process terminates or handles errors appropriately if an API becomes inaccessible.			T.9. Benchmark includes an Oracle solver that can automatically solve all challenges.
Env.	T.4. Residual data or state are fully cleared between runs.			T.10. Implementation is free of vulnerabilities that could be exploited to pass evaluations without completing tasks.
	T.5. Agent is completely isolated from any ground truth information.			
	T.6. Setup does not change over time (e.g., no live website).			

A checklist improves coverage of known failure modes but cannot certify the absence of unknown ones. Adversarial tests and independent checks of accepted outputs remain necessary.

A benchmark score measures a defined construct

State whether the evaluation concerns answer correctness, task completion, human effort, reliable evidence handling or a new scientific result.

A precise estimate of the wrong outcome does not establish the desired capability.
Validate the task and scorer before interpreting a system ranking.

Three evaluation levels

Level	Example unit	Outcome
Model	Question or candidate	Correct answer or valid output
Workflow	Complete trajectory	Completion, cost, interventions
Scientific claim	Result and evidence	Validity, novelty, replication

An evaluation may report several levels. Define each denominator and avoid pooling incomparable outcomes into an unexplained score.

A judge needs a justified evaluation reference

Compare judgments with a justified reference on cases representative of the actual system outputs. Inspect false acceptance and false rejection separately.

Agreement with another model measures agreement, not automatically scientific correctness. A human reference also needs a protocol and uncertainty where judgments legitimately differ.

A synthetic judge audit

Reference label	Judge accepts	Judge rejects
Valid	30	10
Invalid	20	40

There are 100 cases. Acceptance precision is $30/50=0.60$, valid-case recall is $30/40=0.75$, and overall agreement is $70/100=0.70$.

A paired comparison uses discordant outcomes

Operation	Calculation or state	Consequence
Twenty paired tasks	Both correct 7; both wrong 3	Ten concordant tasks
Discordant tasks	Only A correct 8; only B correct 2	Difference is 6/20
Null comparison	$X \sim \text{Binomial}(10, 1/2)$	Condition on discordance
Two-sided exact p	$2 \times (45 + 10 + 1) / 1024 = 0.109375$	No 5% rejection

Synthetic exact McNemar comparison for independent task pairs. Failure to reject does not establish equivalence or rule out a practically important difference.

Worked multiple testing: twenty exploratory analyses

Independent null tests	No false positives	At least one false positive
1	0.950	0.050
5	0.774	0.226
10	0.599	0.401
20	0.358	0.642

Each test has false-positive probability 0.05. Under independence, no false positives has probability 0.95^m ; subtract from one for the family event. The expected number of false positives is a different quantity. Preserve the complete search history before interpreting an agent's most promising reported result.

A simple correction changes the threshold

$$\alpha_{\text{per test}} = \frac{\alpha_{\text{family}}}{m} = \frac{0.05}{20} = 0.0025$$

A Bonferroni rule tests each of twenty hypotheses at $0.05/20 = 0.0025$ to bound the family-wise error rate by 0.05.

The union-bound guarantee does not require independent tests, but the correction can be conservative and may reduce power.

The appropriate analysis depends on the scientific question and dependence structure. A locked confirmation set is another way to separate exploratory selection from a final evaluation.

A small sample leaves substantial uncertainty

Suppose an agent succeeds on 12 of 20 independent, representative tasks. The point estimate is 0.6. A 95% Wilson interval is approximately 0.387 to 0.781.

The interval is wide because the task sample is small. It does not address whether those tasks represent scientific discovery or whether the evaluator is valid.

Statistical uncertainty and construct validity are separate reasons to qualify a broad performance claim.

The structure of a claim–evidence dossier

A claim–evidence dossier connects a bounded scientific claim, the information available before acting, an action and observation trace, and the supporting evidence.

A discovery record connects novelty comparisons, implementation versions, uncertainty, failed attempts and unresolved alternatives. Human contributions define the boundary of the automation claim.

A strong dossier supports exactly the claim it makes. It can establish a useful evaluated capability without presenting every promising research workflow as an autonomous scientist.

A Wilson interval avoids the simplest Wald failure

$$\text{center} = \frac{\hat{p} + z^2/(2n)}{1 + z^2/n}$$

For a binomial success count, the half-width is $z\sqrt{[\hat{p}(1-\hat{p})/n + z^2/(4n^2)]}$ divided by $1+z^2/n$. With 12 successes in 20 trials and $z=1.96$, the interval is approximately [0.387, 0.781].

This models independent trials with a common success probability. Heterogeneous tasks or paired model comparisons require an analysis matching that design.

An interval quantifies sampling uncertainty under assumptions; it does not correct an invalid evaluator.

Confirmation after adaptive search

Under independent null tests at level 0.05, the chance of at least one false positive among twenty tests is about 0.642. The selected smallest p-value cannot be interpreted as if it were the only test.

The Bonferroni threshold $0.05/20$ controls family-wise error without requiring independence, for valid individual p-values and a specified family.

Adaptive campaigns additionally need a justified sequential procedure or fresh confirmation data.

Close the investigation with an auditable claim

Dossier element	Recurring synthetic case	Remaining condition
Claim	Evidence favors temperature-dependent sensor drift	Within the stated candidate model family
Measurement	Units reconciled; reference independently calibrated	Calibration and sampling assumptions remain explicit
Revision history	Expired-reference evidence was retired	Dependent claims were recomputed
Verification	A new measurement probes the rival explanation	The final claim is bounded by its actual support

The scientific output is a supported conclusion plus its provenance and limits. An impressive narrative alone is not the result.

State what would change the conclusion

Possible new evidence	What it challenges	Revision
Fresh reference supports a process change	The sensor-only explanation	Increase support for the process hypothesis
Randomization removes the temperature effect	The proposed temperature mechanism	Reconsider confounding or selection
A third mechanism predicts both observations	Completeness of the candidate family	Expand the model space and test prospectively
Replication reuses the faulty input	Independence of the confirmation	Downgrade the strength of the claim

A precise reversal condition makes an unresolved question intellectually useful. It also separates confidence from resistance to correction.

Recent methods in context

FrontierScience (OpenAI report, 2025/26). Separates Olympiad questions from expert-written scientific research sub-tasks with rubric-based grading. Bounded scientific tasks cover only part of discovery, and scores depend on task selection and the evaluator.

EXP-Bench (ICLR 2026). Evaluates agents on experimental procedures reconstructed from AI research papers and code. Faithful execution of a prescribed experiment is narrower than open-ended creativity or a new scientific discovery.

ResearchBench (ACL Findings 2026). Separates inspiration retrieval, hypothesis composition and ranking in literature-derived scientific tasks. Retrospective reconstruction and a publication-date filter do not establish prospective novelty or eliminate contamination.

References (1/2)

Chenglei Si, Diyi Yang, Tatsunori Hashimoto (2024). Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers

Tarun Gupta, Danish Pruthi (2025). All That Glitters is Not Novel: Plagiarism in AI Generated Research

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun et al. (2025). Establishing Best Practices for Building Rigorous Agentic Benchmarks

Chenglei Si, Zitong Yang, Yejin Choi et al. (2026). Towards Execution-Grounded Automated AI Research

Edwin B. Wilson (1927). Probable Inference, the Law of Succession, and Statistical Inference

Quinn McNemar (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages

References (2/2)

[Samuel Schmidgall, Xiaokai Zhu, Marian Shaw et al. \(2026\). Accelerating Scientific Research with Gemini in the Real-World](#)

[Miles Wang, Robi Lin, Kat Hu et al. \(2025\). FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks](#)

[Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu et al. \(2025\). EXP-Bench: Can AI Conduct AI Research Experiments?](#)

[Yujie Liu, Zonglin Yang, Tong Xie et al. \(2025\). ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition](#)

[Sture Holm \(1979\). A Simple Sequentially Rejective Multiple Test Procedure](#)

[Judea Pearl \(2009\). Causal inference in statistics: An overview](#)