

Hypothesis generation and conceptual novelty

Hypothesis representations and observable predictions.

Reasoning agents, feedback, and scientific evidence

Several stories fit the same observation

Imagine a treatment is followed by a lower measured signal. The observation supports several stories, but repeating those stories in new words adds little.

Ask which commitments an explanation makes: the treatment changes the process, changes the measurement, or coincides with another changing condition.

Changing one commitment at a time creates genuinely different candidate mechanisms. The missing step is constructing predictions that can disagree.

Constructing candidates by changing assumptions

Candidate A changes the underlying process. Candidate B changes the sensor response. Candidate C adds a shared environmental cause.

Each candidate comes from editing a different part of the observation-to-explanation chain. An independent assay, reference sample or environmental control can separate them.

A hypothesis record can now specify entities, context, mechanism and predictions. The record preserves these commitments while later evidence changes their credibility.

Choosing a proposal strategy

Candidate	Reason to choose it	Cost or limitation
Retrieve an established model	A known mechanism fits the setting	May overlook a missing variable or relation
Revise a failed assumption	A residual points to a specific mismatch	Can overfit one observation
Transfer a relation	Another domain supplies useful structure	Mapping and feasibility require checking

Generate candidates from distinct commitments, then select predictions that feasible tests can separate. Novel wording alone is insufficient.

Pedagogical reconstruction

[Wang et al. \(2023\)](#)

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

Generation and selection are different policies

A proposal distribution produces candidates. A selection rule chooses which candidates receive expensive scrutiny or experiments.

The best generator under a broad sampling budget need not be the best generator when only one proposal can be checked. Report proposal counts, filtering rates, and validation costs.

A small set of well-supported ideas can outperform many diverse but infeasible candidates when execution dominates the budget.

Constraints precede scalar preferences

Reject malformed candidates and detect unsupported entities. Check resource feasibility and obvious overlap with prior work. Then compare novelty, plausibility, predicted learning, and cost.

A weighted sum is one decision rule, but its weights encode preferences. Pareto analysis exposes tradeoffs without claiming a universal scientific score.

Keep the rejected candidates and reasons so an evaluator can distinguish creative failure from a broken parser.

Prediction diversity depends on available experiments

Proposal	Prediction in experiment E1	Prediction in experiment E2
A	Positive	Positive
B	Positive	Negative
C	Negative	Positive

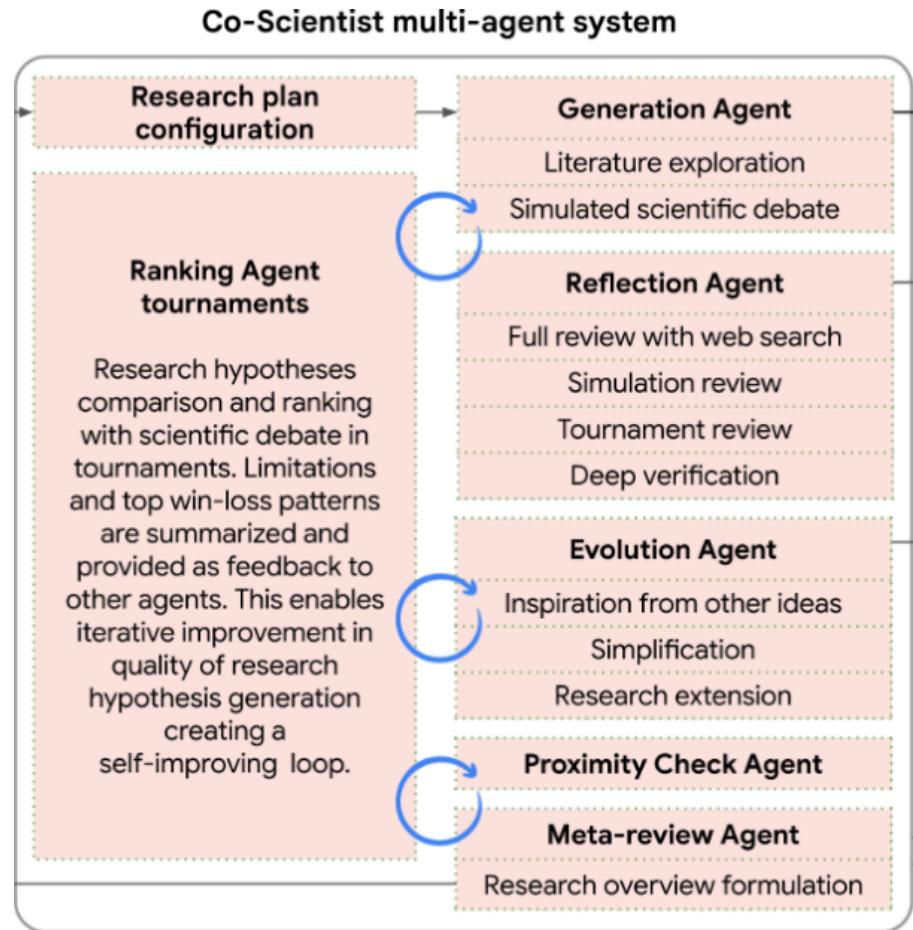
Synthetic predictions. If only E1 is feasible, A and B are experimentally redundant; A and C expose a disagreement.

Feasibility can reverse the preferred pair

Available evidence	Informative pair	Reason
E1 only	A and C	They predict different E1 outcomes
E2 only	A and B	They predict different E2 outcomes
E1 and E2	Any two are distinguishable	Cost and scientific value now matter
Neither experiment	No pair gains evidential diversity	Textual difference alone is insufficient

Novelty, plausibility and testability are distinct. The best portfolio depends on what evidence can actually be obtained.

Co-Scientist: mechanism detail



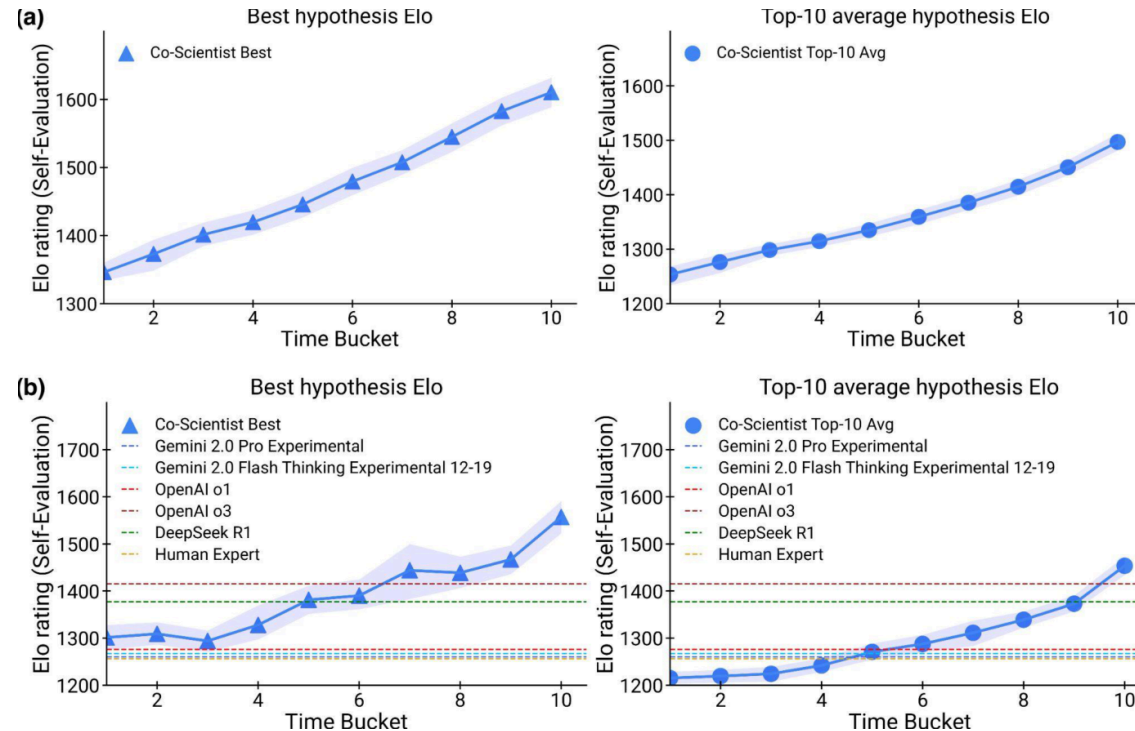
Generation, reflection and ranking form a feedback loop: candidates enter comparisons whose results influence later proposals. A high rank can reflect the evidence available to the evaluator or the persuasiveness of an explanation; the information supplied to each stage determines the interpretation.

Evidence: count support across experimental contexts

Drug combination	MOLM-13	KG-1a
JNJ-64619178 + Selinexor	Synergistic	Mixed response
Palbociclib + Selinexor	Synergistic	Synergistic
JNJ-64619178 + SNDX-5613	Synergistic	Mixed response
CB-839 + Sulfasalazine	Not synergistic	Not synergistic
Venetoclax + Pinometostat	Mixed response	Synergistic
JNJ-64619178 + Selinexor + SNDX-5613	Synergistic	Synergistic
JQ1 + Olaparib + MSA-2	Synergistic	Not synergistic

The source reports seven selected combinations in two cell lines, with biologically independent triplicates. Count only rows labelled “Synergistic” in both columns; “Mixed response” is a separate category. Selection, cell-line context, and the interaction metric all matter. This table records specific laboratory findings, not a clinical success rate.

Co-Scientist: interpretation and limits



These panels show internal Elo and auto-evaluation comparisons. They measure ranking under selected evaluators. The paper also reports experimental validations elsewhere; neither provides a calibrated success rate for every generated hypothesis or general autonomous performance across domains.

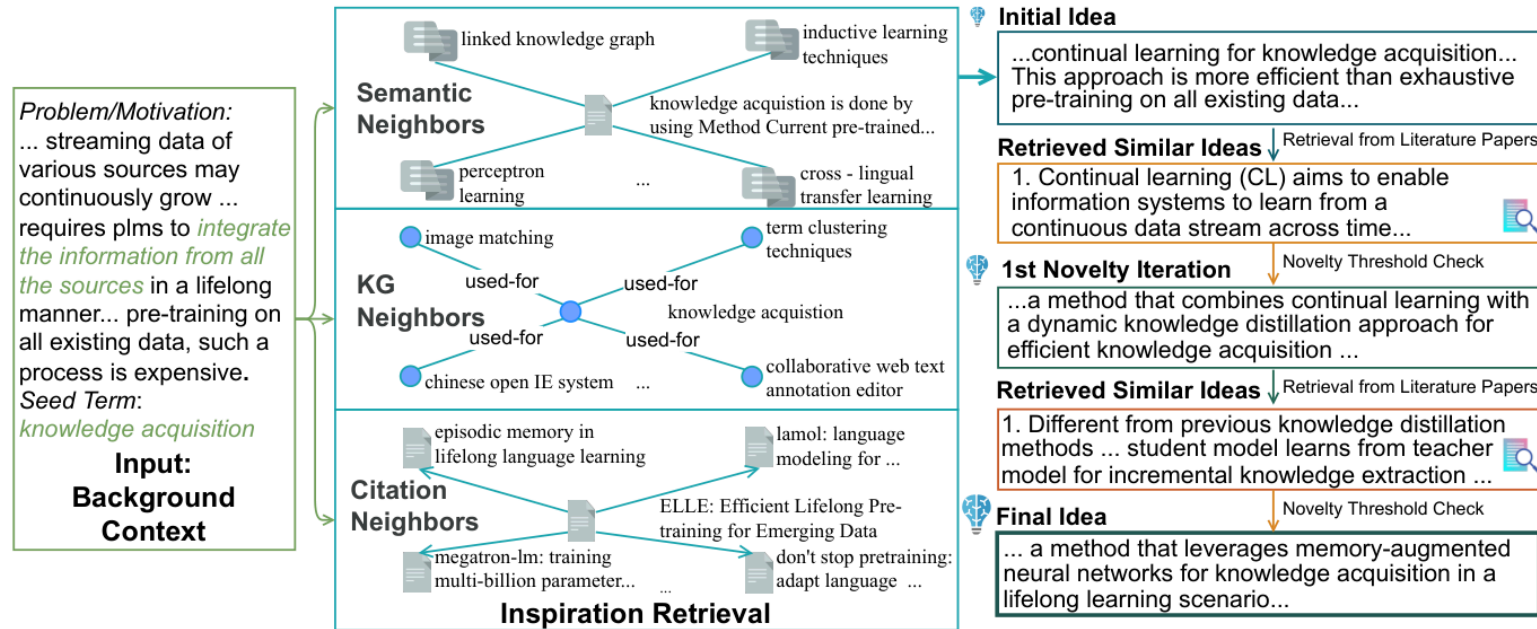
SciMON: the method

SciMON retrieves scientific inspirations from semantic neighbours, knowledge-graph relations, and citation links, then generates a candidate idea from a background problem.

Iterative revision uses a novelty-related comparison against retrieved literature. The design turns retrieval and novelty control into explicit parts of the generator.

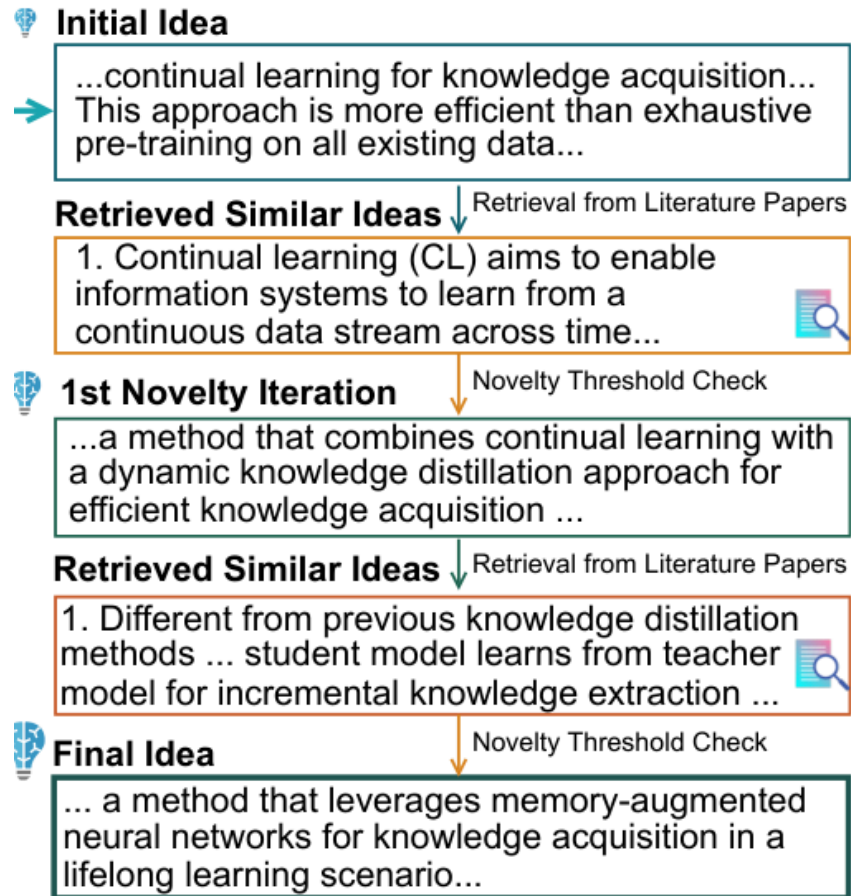
Retrieval quality and the chosen similarity criterion shape which ideas are treated as new, so the novelty objective remains a proxy.

SciMON: retrieval and revision workflow



Background context and three inspiration sources feed iterative idea revision. Retrieval of similar prior ideas can change a proposal. Novel wording and a new scientific relation remain distinct outcomes.

SciMON: mechanism detail



The initial idea is compared with a retrieved predecessor before revision. A changed scientific assumption or relation can support conceptual novelty; different wording alone establishes only textual revision.

Evaluation scope

The study evaluates generated ideas and novelty-oriented revisions under its dataset and human-evaluation protocol.

The table or example compares particular system variants along the stated evaluation dimensions.

Better novelty-related scores indicate improvement under that protocol; they do not independently establish feasibility, conceptual originality against the entire literature, or successful experiments.

SciMON: interpretation and limits

Type	GPT4FS	+SN	+CT	+KG
1st Novelty Δ (%)	+54.4	+55.6	+47.8	+46.7
2nd Novelty Δ (%)	-	+57.8	-	-
1st new terms Δ	+23.1	+22.8	+22.1	+21.9
2nd new terms Δ	-	+21.5	-	-

An incomplete literature index can miss predecessors, while embedding distance can reward superficial changes. Type compatibility, feasibility, and empirical usefulness require additional checks beyond the retrieval objective.

Portfolio selection rewards complementary predictions

Two hypotheses with different wording may predict the same outcomes for every available experiment. They offer little additional decision value when selected together.

Define a distance between predictive distributions under the feasible action set. A portfolio can then favour candidates that expand the range of distinguishable explanations.

This criterion depends on the experiment menu and likelihood models; it is not a general measure of intellectual originality.

Novelty checks have asymmetric errors

A missed prior paper can make a familiar idea look new. An overly broad similarity rule can reject a useful extension as a duplicate.

Retrieve candidate predecessors, compare the central scientific claims, and document the differences. A human expert can adjudicate difficult cases with source evidence.

Evaluate retrieval recall separately from the claim-comparison judgement; otherwise two failure modes are hidden inside a single novelty score.

Execution feedback can change what is selected

An idea may score highly for excitement yet fail because its key experiment is infeasible or its expected effect is absent.

Execution-grounded evaluation adds implementation success, empirical improvement, and robustness to the selection criteria. It also introduces dependence on the executor's competence.

Separate an idea's scientific limitation from failure to implement that idea faithfully. Preserve implementation traces when comparing proposal policies.

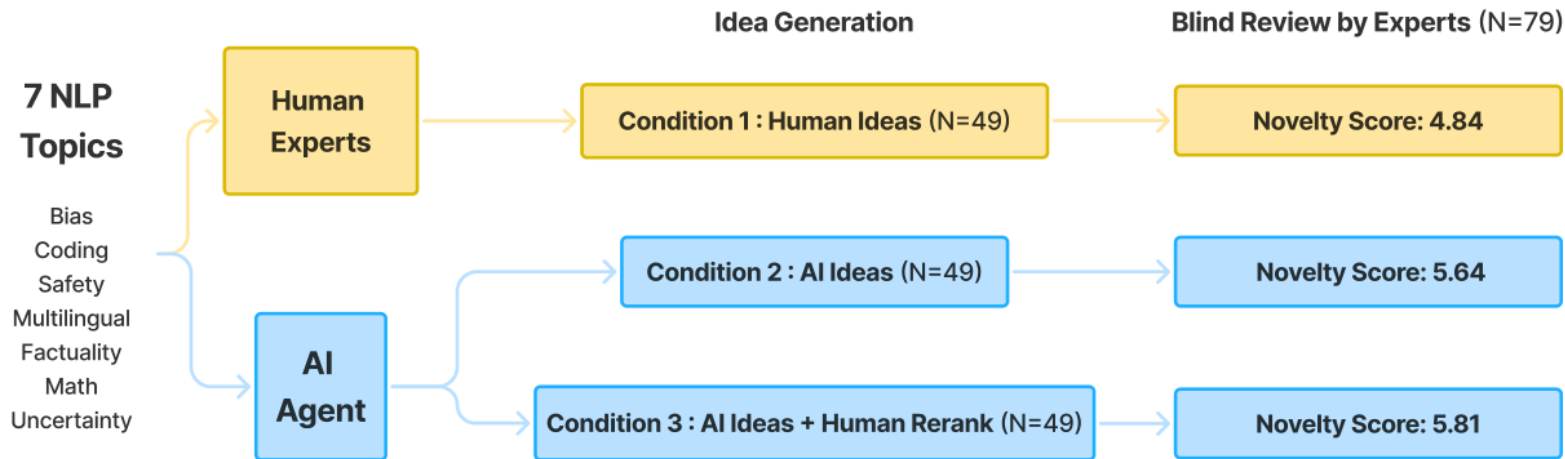
Research ideas: expert evaluation protocol

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.

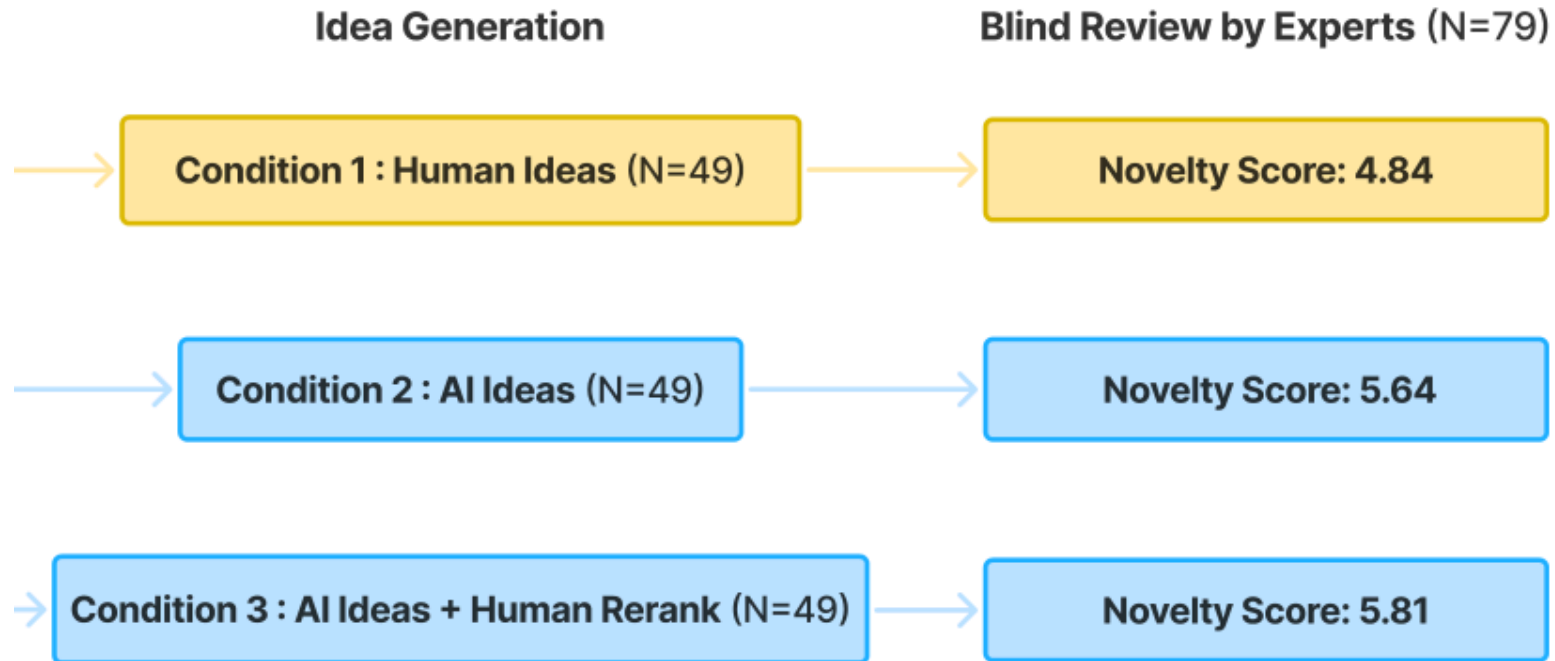
The protocol treats ideation quality as something to measure rather than assuming that a plausible proposal is scientifically valuable.

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.

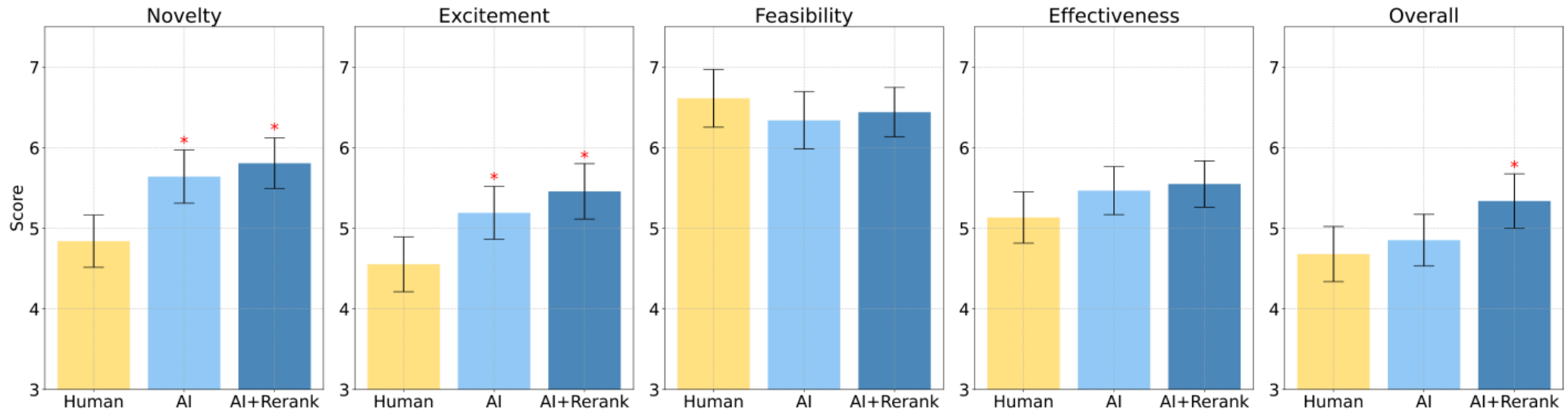
Evaluation scope

The displayed results separate novelty, excitement, feasibility, effectiveness, and overall assessment. AI-generated ideas can score differently on these dimensions.

The evaluation scale and uncertainty intervals constrain the interpretation of differences between conditions.

Expert judgement of an unexecuted idea is informative about perceived quality, but it is not a measurement of the result that its proposed experiment would produce.

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.

Agreement invites a causal question

Observation: binary X and Y always agree. Is changing X enough to change Y, or are both responding to something else?

The conceptual move is to separate prediction from intervention: a pattern can have more than one data-generating explanation.

Construct candidates by changing the parent of Y. Keep the observed pattern fixed so that the new test must do real work.

Build alternatives by changing one equation

Candidate	Structural equations	What is being proposed?
M1: direct effect	U is a fair bit. $X = U; Y = X.$	X transmits the influence to Y.
M2: shared cause	U is a fair bit. $X = U; Y = U.$	U drives both; changing X need not alter Y.
Same observations	$P(X=0, Y=0) = \frac{1}{2}$ $P(X=1, Y=1) = \frac{1}{2}$	Verbal differences become executable commitments.

Neither model wins by fitting these observations better. A new kind of evidence is needed.

Choose a test where predictions separate

Action	Prediction under M1	Prediction under M2
Observe $X = 1$	$Y = 1$	$Y = 1$
Set $X = 1$	$Y = 1$ with probability 1	$Y = 1$ with probability $\frac{1}{2}$
Selection principle	Change X's equation; retain $Y = X$.	Change X's equation; retain $Y = U$.

The useful alternative is one whose consequences differ under a feasible, valid intervention.

An executable hypothesis is more than a label

A candidate specifies variables, structural assumptions and observable predictions. A fluent explanation without these commitments cannot face this test.

A predictor may be enough for forecasting; a causal model is needed for this intervention question. More complex models remain credible if the simple pair fails.

Carry the same case forward: L12 checks identification; L17 uses experimental outcomes to revise belief without pretending the candidate set is complete.

Portfolio selection uses marginal coverage gains

Operation	Calculation or state	Consequence
Weights	a:3, b:2, c:1, d:4	Value of covered features
Candidates	$A=\{a,b\}$; $B=\{b,c\}$; $C=\{d\}$	Coverage can overlap
First selection	A:5, B:3, C:4	Choose A
Second selection	$\text{Gain}(B A)=1$; $\text{gain}(C A)=4$	Choose C; total value 9

Synthetic weighted coverage objective with a budget of two candidates. A feature represents a defined scientific prediction or capability, rather than a word in the hypothesis.

Worked portfolio: three candidates, two slots

A public simulation task offers candidates A, B, and C. Their synthetic utility estimates are 8, 7, and 6. The budget allows two candidates.

A and B rely on the same mechanism; C tests a different one. Let the redundancy penalty be five for the pair A–B and zero for the other pairs.

This simple setting makes the difference between ranking individual ideas and selecting a complementary portfolio explicit.

Calculate the best portfolio

Portfolio	Utility sum	Redundancy	Net score
A + B	$8 + 7 = 15$	5	10
A + C	$8 + 6 = 14$	0	14
B + C	$7 + 6 = 13$	0	13

Select a set, not merely the two largest individual scores. A and B share a mechanism; C adds a distinct testable prediction. The penalty scale determines this ranking and needs justification. The following feasibility constraint changes the candidate set without making C scientifically false.

Convert diversity into experimental predictions

Suppose A and B both predict an increase after perturbation e_1 , while C predicts a decrease. Selecting A and C yields a sharper contrast for this experiment.

If the available experiments cannot distinguish A from C either, the stated diversity reward has little operational value.

Portfolio quality is therefore conditional on the feasible measurement set. Re-evaluate it when a new assay, dataset, or simulator becomes available.

Sensitivity analysis is part of the answer

$$U(\{A, B\}) = 15 - r, \quad U(\{A, C\}) = 14$$

Let the A–B redundancy penalty vary from zero to five. The score of A–B becomes 15 minus the penalty; A–C remains 14 in the preceding example.

A–B is preferred below a penalty of one, the pairs tie at one, and A–C is preferred above one.

The decision boundary describes whether the selected portfolio is robust or largely determined by an uncertain preference parameter.

Complementarity as a marginal portfolio gain

$$\Delta(i | S) = u_i - \lambda \sum_{j \in S} d_{ij}$$

For a pairwise redundancy penalty, adding candidate i to portfolio S contributes its utility u_i minus its overlap penalties with existing members. The score uses unordered pairs once.

A candidate with lower standalone utility can have higher marginal value if it adds a distinct prediction.

Greedy selection is a heuristic unless additional properties and constraints justify a guarantee. Negative marginal gains can make filling every available slot undesirable.

A sensitivity analysis for the hypothesis portfolio

In the worked example, the scores of AB and AC are $15-\lambda$ and 14 when only AB has unit redundancy. AC is preferred for $\lambda > 1$ under this parameterisation.

The earlier penalty of five corresponds to $\lambda=5$. Different conventions for the overlap matrix must not silently change the objective.

Changes in the selected portfolio across scientific preference weights expose sensitivity that a single fixed-weight ranking conceals.

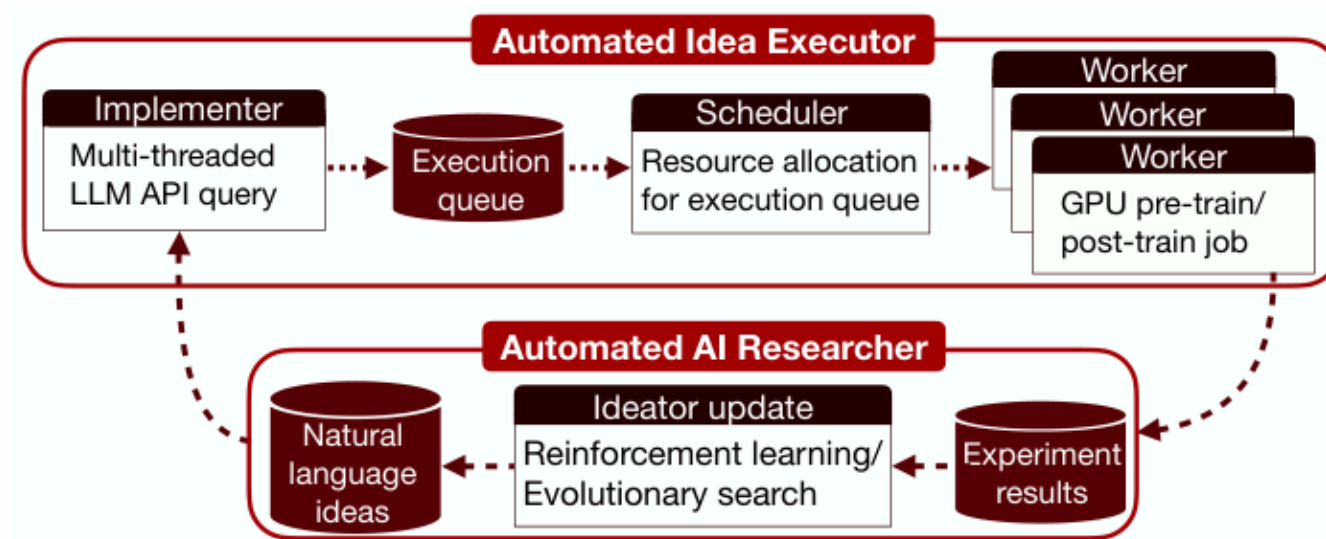
Recent research: Execution-grounded AI research

The study separates an ideator from an automated executor containing implementation, scheduling, and worker components.

Proposed ideas are implemented and tested in defined machine-learning environments. Execution feedback can then guide evolutionary search or reinforcement learning of the ideator.

The literature snapshot is dated 11 September 2026. Findings remain conditional on the source version and evaluation conditions. Recent preprints contribute reported results with distinct validation requirements.

Execution-grounded research: feedback architecture



The architecture links a natural-language proposal to implementation, an execution queue and worker results, then returns feedback to the researcher. It explicitly separates proposing an idea from obtaining an executed result.

Recent methods in context

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.

Co-Scientist (Nature 2026). Co-Scientist organizes hypothesis generation, critique and tournament-based refinement. The Nature publication adds peer-reviewed evidence for selected applications. Internal rankings and external scientific validation support different claims.

CHIVE (Aug 2026 preprint). Investigates observed model behaviors with counterfactual prompt edits and tests explanations through predictive usefulness. The reported lack of uplift from tested interpretability tools is conditional on this evaluation and tool set.

References (1/2)

[Juraj Gottweis, Wei-Hung Weng, Alexander Daryin et al. \(2025\). Accelerating scientific discovery with Co-Scientist](#)

[Qingyun Wang, Doug Downey, Heng Ji et al. \(2023\). SciMON: Scientific Inspiration Machines Optimized for Novelty](#)

[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](#)

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

References (2/2)

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

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

[Adam Karvonen, Euan Ong, Subhash Kantamneni et al. \(2026\). Would this change your answer? Evaluating Explanations of LLM Behavior In The Wild with Counterfactual Experiments](#)

[Kathryn Chaloner, Isabella Verdinelli \(1995\). Bayesian Experimental Design: A Review](#)