

Scientific discovery as a computational process

Discovery as a partially observed decision process.

Reasoning agents, feedback, and scientific evidence

A result that changes the next question

Imagine a research assistant proposes that a sensor drift reflects a new physical effect. A fresh calibration could instead overturn the premise.

The next useful action depends on what that calibration reveals. A single answer cannot specify the whole investigation in advance.

This makes the sequence of questions, observations and revisions worth modeling. The opportunity is to improve how evidence changes the next decision.

The research process becomes the object

Separate the event we observed from the explanation we currently favor. A correction may change the explanation without erasing the historical observation.

Now give the system a record of evidence, candidate explanations and remaining resources. A policy can choose the next action from that record.

This construction suggests an adaptive research loop. Whether a generated hypothesis, algorithm or empirical result counts as discovery still requires its own checks.

Choosing an investigation architecture

Candidate	Reason to choose it	Cost or limitation
One-shot proposal	A cheap candidate for independent testing	Cannot react to evidence arriving later
Fixed workflow	Known steps and stable prerequisites	Unexpected observations need a new branch
Adaptive investigation	Observations can change the next action	Needs reliable state, tools and checking

Choose adaptation when new evidence can change the plan enough to justify its cost. A fixed workflow remains useful for stable procedures.

Pedagogical reconstruction

[Kaelbling et al. \(1998\)](#)

[Lu et al. \(2024\)](#)

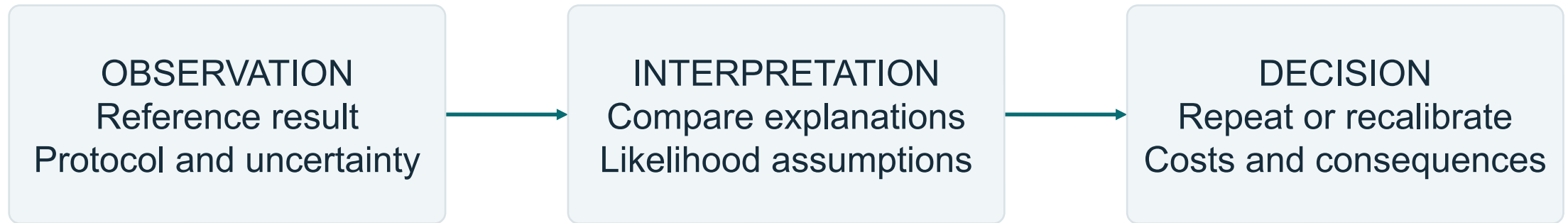
A belief state makes uncertainty explicit

$$b_t(h) = p(h \mid D_t), \quad a_t \sim \pi(a \mid b_t, \text{budget}_t)$$

Let h denote a candidate mechanism, D the observations collected so far, and $b(h)$ the current belief distribution. An action can acquire data, run a computation, or propose another candidate.

The policy chooses actions using the belief state and remaining resources. The state is a modelling choice: structured records make this choice inspectable, but natural-language representations can also support useful inference.

An observation is not its interpretation



A reliable observation can support several interpretations. An interpretation can justify different actions under different losses. Preserve the three objects and their links: a corrected measurement may change the inference, while a changed objective can change the decision even when the evidence stays fixed.

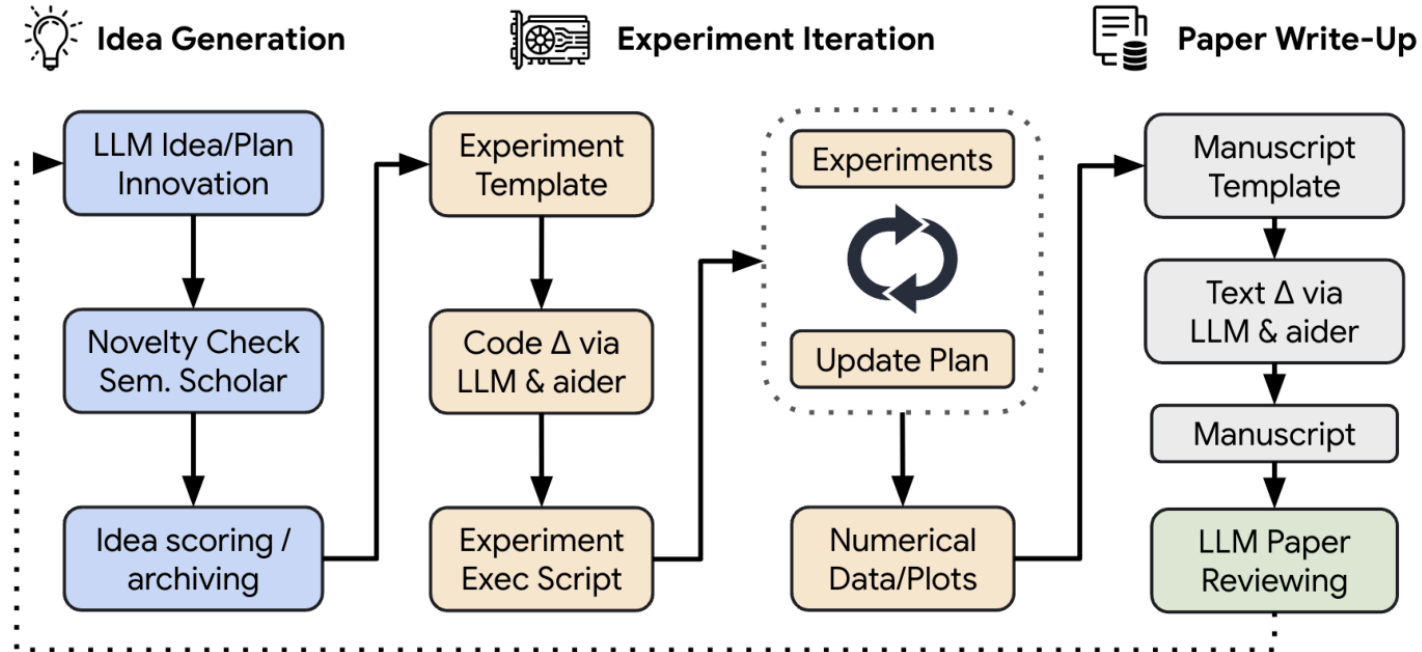
The AI Scientist: the method

An LLM proposes research ideas from a supplied experimental setting, checks nearby literature, modifies code, runs experiments, and assembles a manuscript.

Numerical outputs and plots inform subsequent iterations. An automated reviewer provides an additional assessment of the resulting paper.

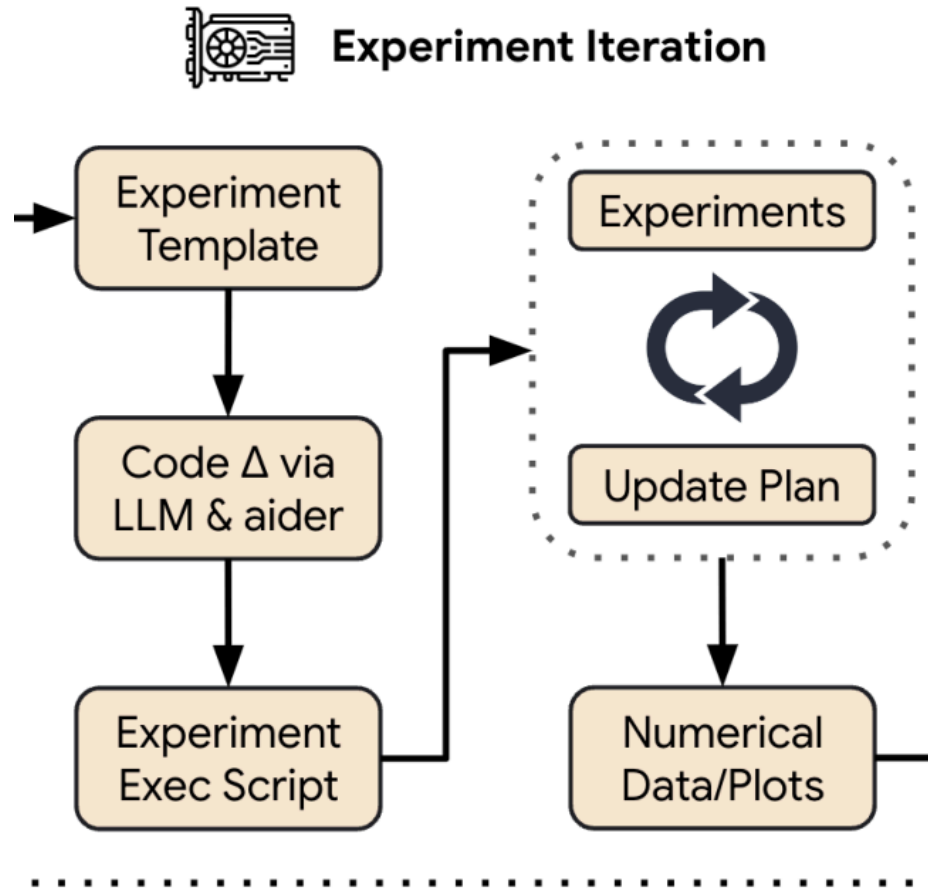
The workflow combines scientific assistance with orchestration and execution; the quality of each stage constrains the usefulness of the complete output.

The AI Scientist: research workflow



The three columns connect idea generation, experiment iteration and writing. Measurements and generated text travel through different parts of this loop. The novelty check and paper reviewer judge different targets; neither is an independent experimental replication.

The AI Scientist: mechanism detail



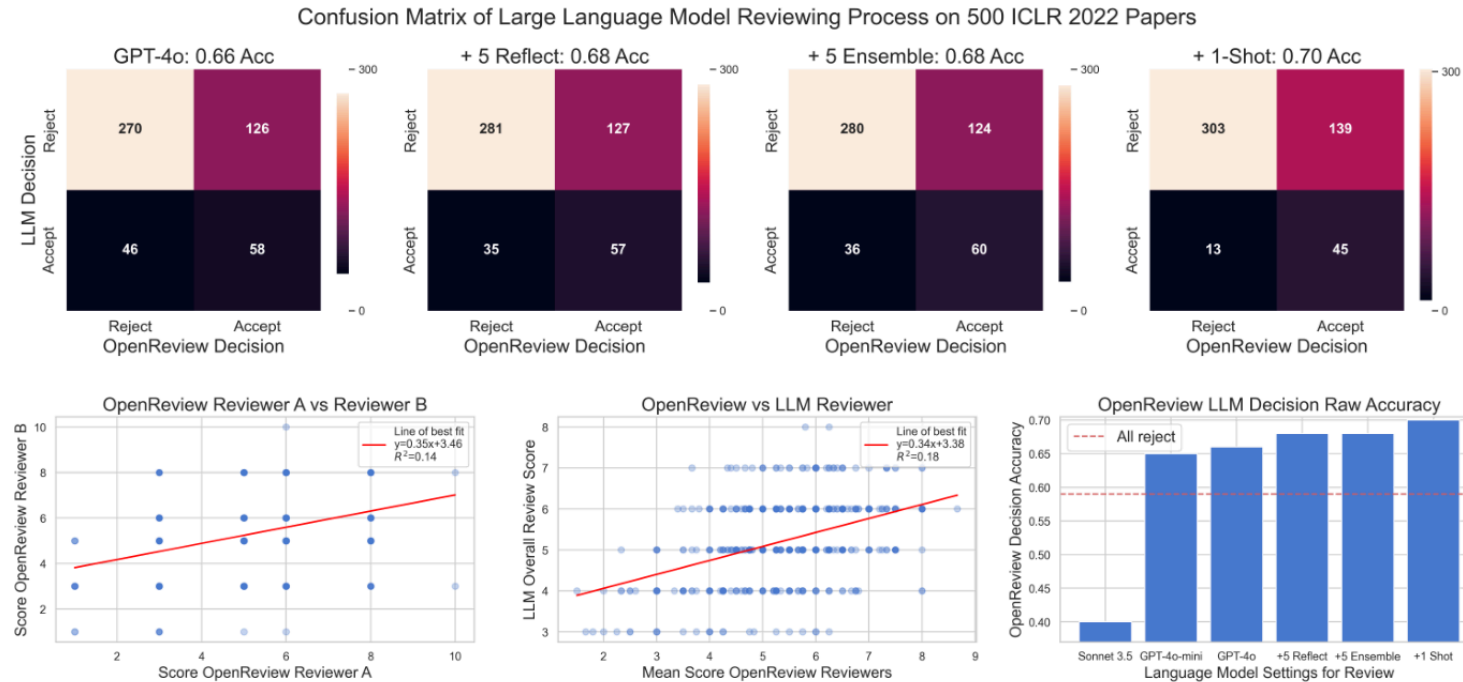
In the experiment loop, code changes produce numerical data and plots that feed an updated plan. Execution records distinguish changes supported by new observations from revisions to the written report.

Evidence: reconstruct the reviewer metric

	Reviewer	Balanced Acc. ↑	Accuracy ↑	F1 Score ↑	AUC ↑	FPR ↓	FNR ↓
	Human (NeurIPS)¹	<u>0.66</u>	<u>0.73</u>	<u>0.49</u>	<u>0.65</u>	<u>0.17</u>	<u>0.52</u>
	Random Decision	0.50	0.50	0.40	0.50	0.50	0.50
	Always Reject	0.50	0.59	0.00	0.50	0.00	1.00
Uncalibrated	Sonnet 3.5	0.52 ± 0.01	0.40 ± 0.01	0.55 ± 0.01	0.52 ± 0.01	0.95 ± 0.02	0.00 ± 0.00
	GPT-4o-mini	0.53 ± 0.02	0.65 ± 0.01	0.11 ± 0.06	0.53 ± 0.02	0.01 ± 0.01	0.94 ± 0.04
	GPT-4o (0-shot)	0.61 ± 0.04	0.68 ± 0.03	0.43 ± 0.07	0.61 ± 0.04	0.11 ± 0.03	0.67 ± 0.07
	GPT-4o (1-shot)	0.60 ± 0.03	<u>0.70 ± 0.03</u>	0.37 ± 0.08	0.60 ± 0.03	0.04 ± 0.02	0.76 ± 0.06
Calibrated	Sonnet 3.5 @8	0.59 ± 0.04	0.65 ± 0.04	0.45 ± 0.06	0.59 ± 0.04	0.20 ± 0.04	0.61 ± 0.07
	GPT-4o-mini @6	0.59 ± 0.04	0.64 ± 0.04	0.45 ± 0.06	0.59 ± 0.04	0.22 ± 0.05	0.60 ± 0.07
	GPT-4o (0-shot) @6	0.63 ± 0.04	0.63 ± 0.04	0.56 ± 0.05	0.63 ± 0.04	0.38 ± 0.05	0.36 ± 0.07
	GPT-4o (1-shot) @6	<u>0.65 ± 0.04</u>	<u>0.66 ± 0.04</u>	<u>0.57 ± 0.05</u>	<u>0.65 ± 0.04</u>	<u>0.31 ± 0.05</u>	<u>0.39 ± 0.07</u>

The calibrated one-shot GPT-4o row reports false-positive rate 0.31, false-negative rate 0.39, and balanced accuracy 0.65. Reconstruct the last number. These AI results use 500 ICLR papers; the human row comes from a separate NeurIPS consistency study. The table does not establish a matched human–AI comparison.

The AI Scientist: interpretation and limits



The experimental domains and templates restrict the action space. Novelty screening and automated review remain fallible, and successful code execution does not validate every inference in a manuscript.

Follow one investigation, not one answer

State	Action or observation	What changes?
A: sensor drift; B: process change	Compare with an independent reference	A test can separate explanations
Prior: $P(A) = P(B) = 0.5$	A positive calibration check arrives	The evidence favors A
$P(+ A) = 0.8$; $P(+ B) = 0.2$	Bayes update: $P(A +) = 0.8$	Confidence rises; certainty does not
Reference validity	Validate the reference and units	The measurement model can also fail

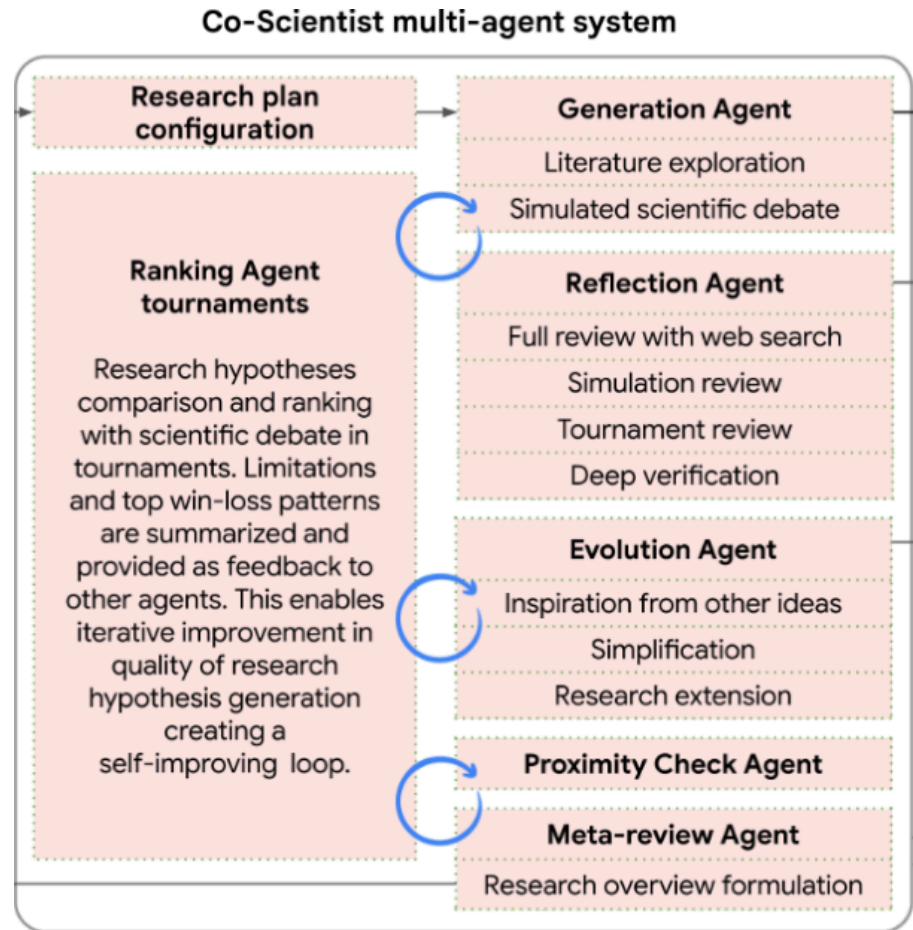
Synthetic case used throughout the series. A scientific agent must preserve both its conclusion and the assumptions supporting it.

Agent state and scientific belief are different

Object	Typical representation	What it does not establish
Proposal	Text, code or a candidate object	That the candidate is novel or correct
Search priority	A score or tournament ranking	A calibrated probability of truth
Observation	A tool result with units and provenance	An independent replication
Scientific belief	A model with explicit assumptions	That the model family contains the truth

Co-Scientist uses generation, reflection and ranking. Bayesian belief here is an analytical lens, not a claim about its implementation.

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.

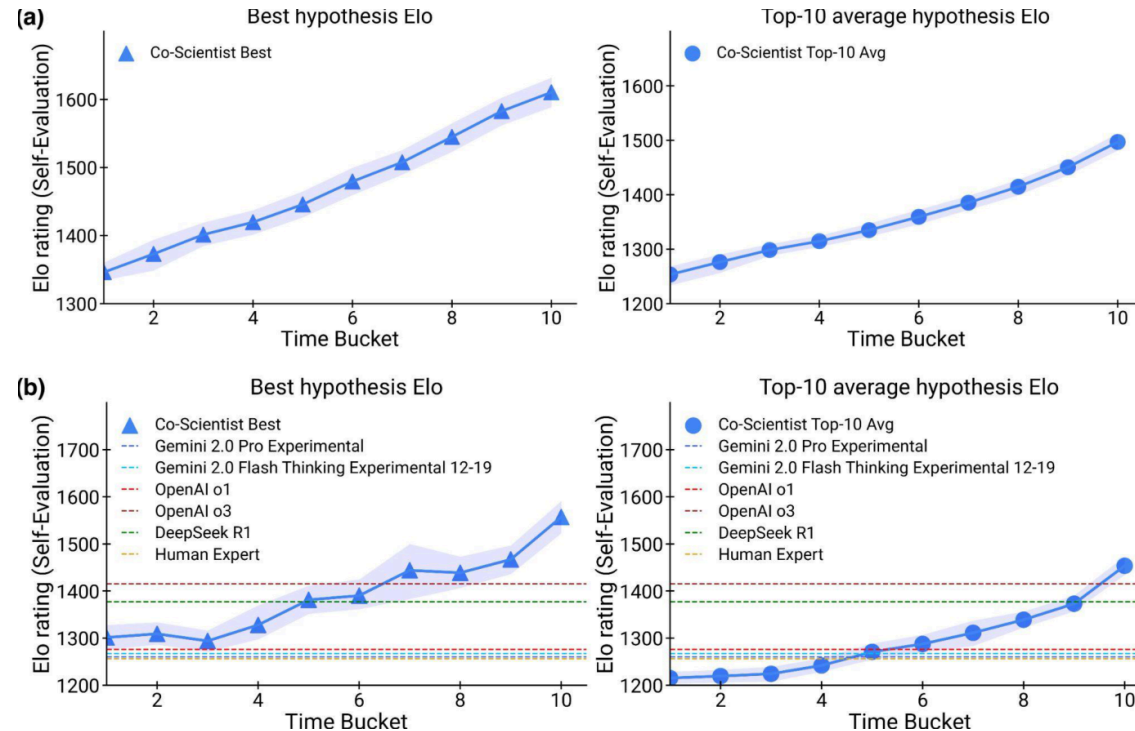
Evaluation scope

The source compares hypothesis quality as inference resources increase, using internal Elo evaluations and additional expert assessments.

The panels use different evaluators and denominators. Automatic rankings, human judgments and experimental outcomes support different conclusions.

Improvements in tournament score support a claim about the ranking process; independent validation is needed to connect that signal to scientific utility.

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.

Verification is a stack of different checks

A code test checks an implementation property. A held-out evaluation checks predictive performance. A replication checks stability under repeated procedures. A novelty search checks overlap with known work.

A language-model critic can help allocate effort at any stage, but its judgement is fallible and can share errors with the generator.

Independent evidence matters more than simply assigning the reviewer a different model name.

The objective determines the architecture

Information gain values learning; expected improvement values finding a better design; decision utility values consequences of acting. A single experiment can rank differently under these objectives.

A budget constrains total expenditure and delays, not just the next action. An information-per-cost ratio is a useful myopic heuristic, but it need not solve a multistep allocation problem.

The scientific objective determines which search method is appropriate.

A minimal, replayable research loop

Propose candidates with explicit predictions. Select an action using an objective and the available resources. Execute it and preserve the original result.

Validate the measurement, update beliefs, and record the reason for the update. Decide whether to continue, change the model class, or report unresolved uncertainty.

The loop is an engineering abstraction. It supports discovery work without claiming that all scientific creativity reduces to one algorithm.

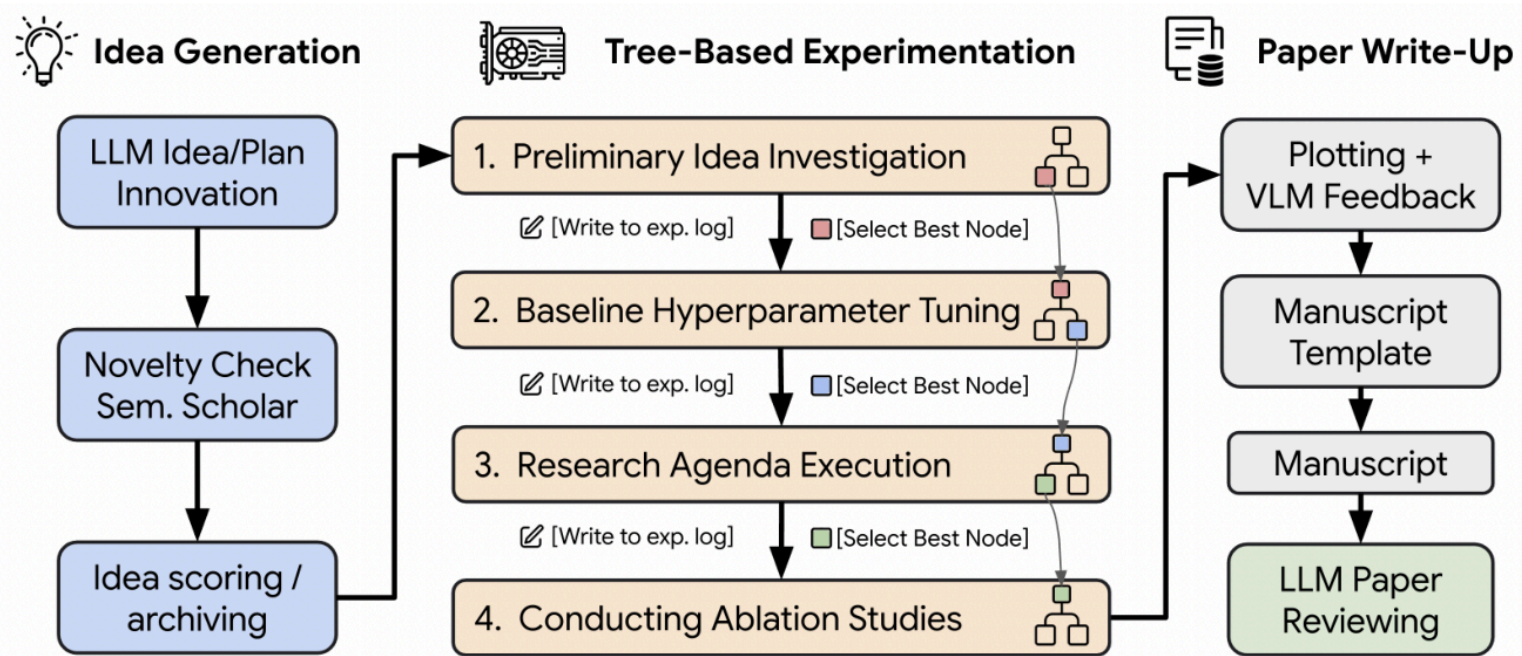
The AI Scientist-v2: the method

The workflow uses agentic tree search to organise experimental research rather than relying exclusively on a fixed experiment template.

Research stages include preliminary investigation, tuning, agenda execution, and ablation, followed by figure creation and writing. Candidate nodes preserve alternative computational paths.

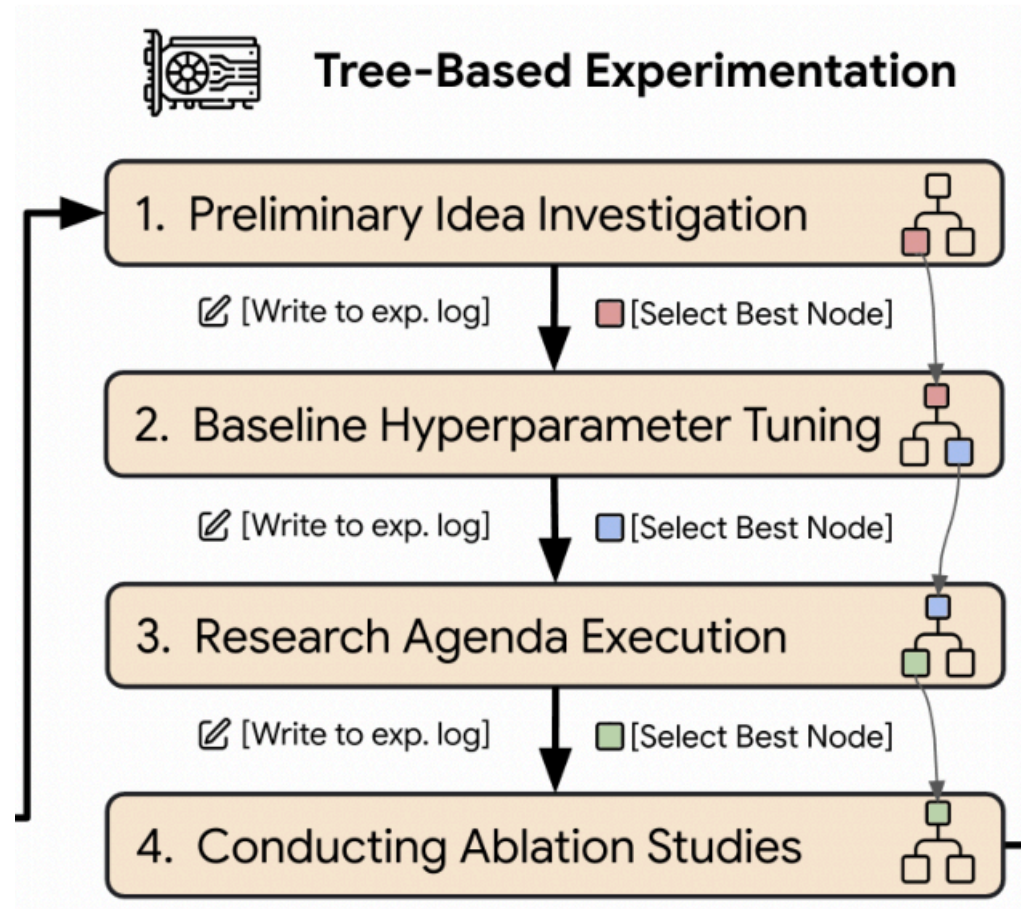
An evaluator helps choose promising branches, but the scientific interpretation of a branch still depends on its implementation and empirical evidence.

The AI Scientist-v2: staged search workflow



The four experimental stages retain selected nodes and pass their records to writing. The tree organizes alternatives and attempts. Scientific likelihood and novelty require evidence beyond the search structure.

The AI Scientist-v2: mechanism detail



The transition from investigation to tuning selects a retained node. Its evidential basis and the treatment of discarded branches determine whether the resulting search trace supports the selection.

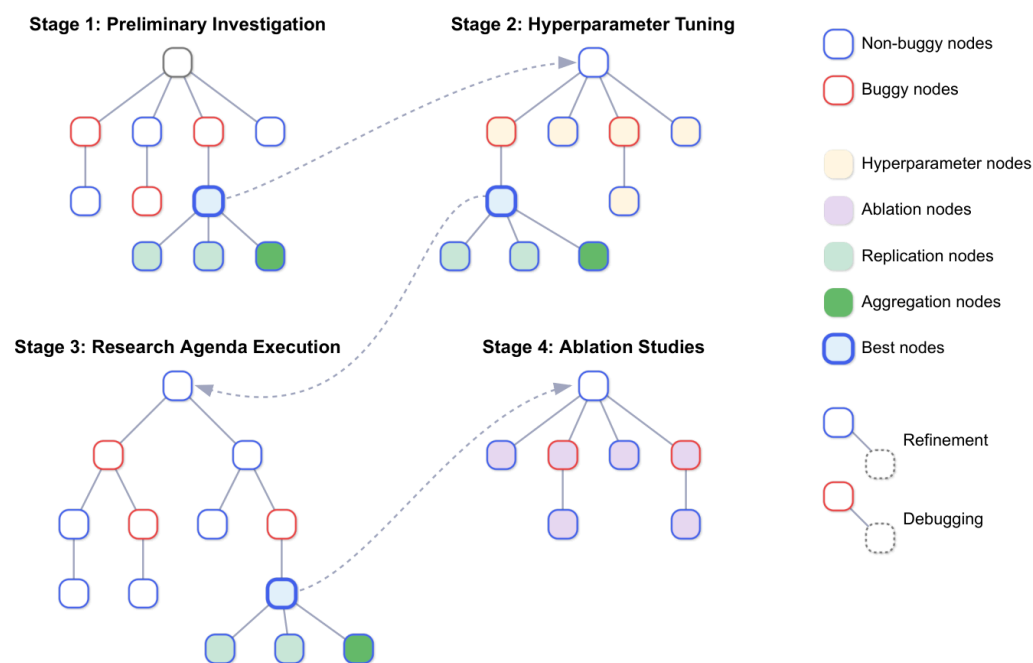
Evaluation scope

The source presents workshop-level generated research and a tree-structured experimental procedure. The extracted search-stage figure shows how node types and selected paths differ.

It is an architectural illustration, not a plot of causal benefit from tree search.

To estimate that benefit, compare a matched sequential workflow using the same tools, evaluation procedure, and total experimental budget.

The AI Scientist-v2: interpretation and limits



Buggy and non-buggy branches lead to a selected checkpoint for the next stage. Debugging success and favorable model evaluation concern different targets. The workflow describes search organization rather than establishing novelty.

The course question

Reasoning agents use feedback to revise decisions. A scientific claim additionally requires evidence appropriate to its scope.

We will distinguish the proposed object, the mechanism that changes it, and the evidence that licenses a conclusion. A better search score can justify further testing without establishing a discovery.

Discovery objects require different checks

Output	Immediate check	Further scientific obligation
Hypothesis	Specific prediction	Discriminating evidence
Algorithm	Specification and cost	Generality and prior art
Empirical finding	Measurement validity	Replication and relevance
Mechanism	Identified causal model	Alternative explanations

These are analytical categories. One research trajectory can produce several outputs, each with its own evidential requirement.

A claim has a population and a protocol

Write a result as a claim about an outcome, a population of tasks and a specified information/resource regime. Preserve the version of the system and evaluator.

Changing retrieval access or allowing expert intervention changes the regime. The same accuracy number then describes a different scientific comparison.

A short research trace

Event	Record	Permitted update
Candidate proposed	Mechanism A predicts drift	Add a hypothesis
Sensor check fails	Calibration invalid	Repair execution
Independent measurement	Valid drift estimate	Update mechanism belief
Literature match found	Prior result with same scope	Revise novelty claim

The record changes at every event. The scientific belief changes only when the event carries relevant evidence under the model.

A belief update changes the optimal action

Operation	Calculation or state	Consequence
Prior	$P(A)=0.6, P(B)=0.4$	Choose A: value 0.6
Positive test	$P(+)=0.55; P(A +)=9/11$	Choose A after +
Negative test	$P(A -)=1/3$	Choose B after -
Test then act	$0.45+0.30-0.05=0.70$	Testing beats stopping

Synthetic symmetric test: $P(+|A)=0.75$ and $P(+|B)=0.25$. A correct terminal choice pays one unit. The test costs 0.05 units.

Worked case: two explanations of a sensor drift

Candidate explanation	Prior	P(positive hypothesis)
A: temperature-dependent calibration error	0.5	0.8
B: change in the monitored process	0.5	0.2

A sensor drifts during a long campaign. A controlled reference measurement discriminates two candidate explanations. These synthetic likelihoods assume the reference procedure is valid; they are not measurements from a named instrument.

Predict the outcome before observing it

$$p(+) = 0.5 \times 0.8 + 0.5 \times 0.2 = 0.5$$

Marginalise over the competing explanations. Under equal priors, a positive reference result has probability 0.5.

This predictive probability must be computed before seeing the result. It is the weighting distribution used by an experiment-selection calculation.

A model that can only write a retrospective explanation has not supplied a likelihood for prospective design. An approximate or learned likelihood can still be useful if its uncertainty is evaluated.

Posterior updates and remaining uncertainty

$$p(A \mid +) = \frac{0.8 \times 0.5}{0.5} = 0.8$$

Suppose the reference result is positive. Bayes' rule gives posterior probability 0.8 for calibration error. The competing explanation retains probability 0.2.

The observation favours A; it does not prove A. A second test may be valuable, especially if its errors differ from the first test.

A reproducible update records the prior, likelihood assumptions, observation identifier and posterior. These records allow another researcher to recompute the update when an assumption changes.

A finite scientific decision problem

A finite scientific decision problem consists of an explicit candidate set and actions with distinct predictions and costs. This structure generalises the two-hypothesis example.

A research trace separates the pre-action belief, chosen action, observed result and resulting belief. A failed measurement can leave the scientific belief unchanged while revealing an execution problem.

The comparison is useful when its decisions and limitations are replayable. It need not claim a previously unknown scientific result.

From one experiment to a finite-horizon policy

$$V_t(b) = \max_e \left\{ -c(e) + \sum_y p(y|b, e) V_{t+1}(b_{e,y}) \right\}$$

Here b is a belief over hypotheses, $c(e)$ is action cost, and $b_{e,y}$ is the Bayesian update after outcome y . The terminal value represents the scientific decision objective.

The equation assumes a finite action and outcome set and a known observation model. A stop action returns the value of deciding now.

The sensor example supplies one update inside this policy; it does not solve the entire sequential decision problem.

Decision policies under a shared likelihood model

For the sensor case, a fixed test sequence, a greedy information policy and a two-step lookahead policy differ in their use of future observations under the same likelihood model.

A controlled comparison fixes candidate mechanisms and total measurement budget while varying how future outcomes influence action choice.

Decision error, measurement cost and test-selection frequencies distinguish the benefit of adaptive planning from changes in proposal generation or measurement quality.

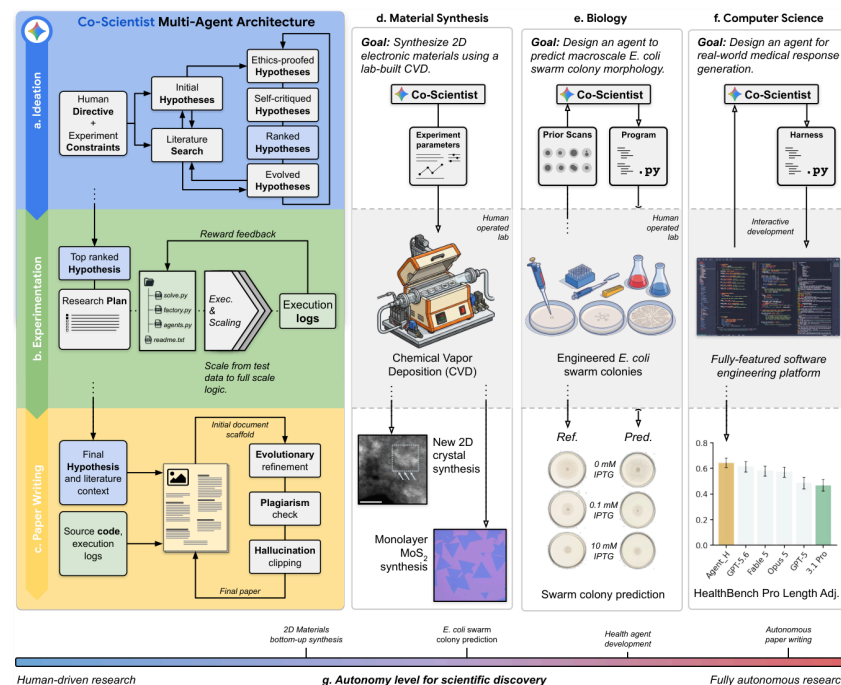
Recent research: Co-Scientist in real-world research

This August 2026 study extends Co-Scientist across materials science, biology, and computer science, with different degrees of physical execution and human involvement.

It reports laboratory-linked workflows and evaluations of generated research, enabling a more concrete discussion of closed-loop operation than a proposal-only demonstration.

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.

Co-Scientist: hypothesis-to-execution workflow



The workflow connects hypothesis generation, execution and reporting. External measurements can ground the loop, while the individual experiments determine the strength of the validation.

Recent methods in context

Aletheia (preprint; revised 2026). Uses generation, verification and revision in natural language for research-level mathematics. The strength of each discovery claim depends on human contributions, priority searches and independent mathematical checking.

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.

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.

References (1/2)

[Chris Lu, Cong Lu, Robert Tjarko Lange et al. \(2024\). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery](#)

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

[Yutaro Yamada, Robert Tjarko Lange, Cong Lu et al. \(2025\). The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search](#)

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

[Leslie Pack Kaelbling, Michael L. Littman, Anthony R. Cassandra \(1998\). Planning and acting in partially observable stochastic domains](#)

References (2/2)

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

[Tony Feng, Trieu H. Trinh, Garrett Bingham et al. \(2026\). Towards Autonomous Mathematics Research](#)

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

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

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