

DiscoveryRSI: Recursively Learning to Discover Valuable Hypotheses

Jianyuan Zhong^{1,2,*} Chenwei Wu^{3,*} Zhaochen Yu^{4,*} Xixian Jiang¹¹ Jiajun Fan¹⁴ Xuewei Li⁶ Hao Xiao⁵
Xilin Dai² Junyang Wang⁷ Jiaru Zou⁸ Zhiqing Cui⁹ Boyuan Zheng³ Jing Xiong¹⁰ Zhen Zhang¹⁰
Qiang Gao⁵ Michael Morley¹² Luis Nakayama^{13,14} Qiang Xu² Liyue Shen³ Tao Gui⁵
Leo Anthony Celi^{12,14,†} James Zou^{8,†} Zhenfei Yin^{1,15,†} Yingcheng Wu^{1,8,†} Ling Yang^{1,7,16,†}

¹PhAI Labs ²The Chinese University of Hong Kong ³University of Michigan

⁴National University of Singapore ⁵Fudan University ⁶Peking University People's Hospital

⁷Peking University ⁸Stanford University ⁹HKUST ¹⁰University of California, Santa Barbara

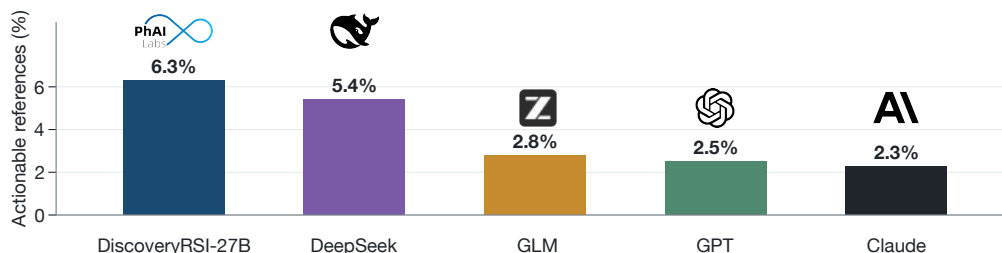
¹¹Southeast University ¹²Harvard Medical School ¹³Federal University of São Paulo ¹⁴UIUC

¹⁵Massachusetts Institute of Technology ¹⁶University of Oxford ¹⁷Princeton University

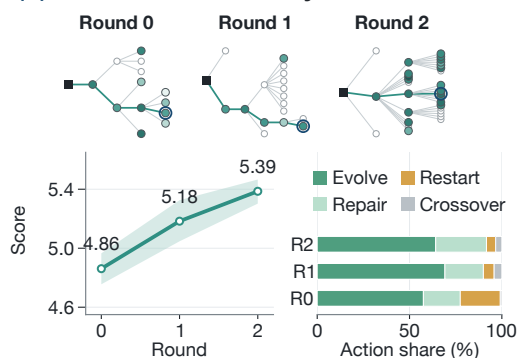
*Equal contribution †Corresponding authors

[Code: GitHub](#) | [Models: Hugging Face](#) | [Website: DiscoveryRSI Arena](#)

(a) Hypothesis hits in subsequent published research



(b) RSI shifts evolutionary behavior



(c) Greater gains per revision

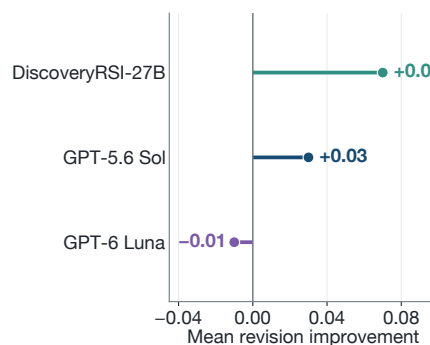


Figure 1 | (a) More actionable links to published research. The rate is the fraction of generated hypotheses linked to post-cutoff expert publications supplying a concrete design change, comparator, measurement, or testable evidence gap. DiscoveryRSI leads at 6.3% (Appendix F). **(b) RSI improves quality and shifts evolutionary behavior.** Score measures research quality and action share is the fraction of episodes using each operation. Research quality rises from 4.86 to 5.39 as evolve/repair use rises from 77.5% to 91.7%. **(c) Larger human-rated gains per revision.** Mean revision change averages the human-score difference between a hypothesis and its predecessor: +0.07 for DiscoveryRSI, +0.03 for Sol, and -0.01 for Luna.

Abstract

Scientific discovery requires developing worthwhile hypotheses into practical investigations, yet learning from this evolving research process remains challenging. We introduce an operational definition of **valuable and testable research hypotheses**: evidence-grounded claims about consequential uncertainties, coupled with feasible investigations, explicit decision criteria, and alternative explanations. Our **DiscoveryRSI** framework develops hypothesis trees through multi-agent evolutionary search, compiles selected investigations into executable Harbor tasks, and recursively learns through three mechanisms: **hypothesis credit optimization** attributes scientific assessments to documented research contributions; **harness replay** identifies productive strategies and recurring failures to refine research workflows; and **re-discovery** applies accepted improvements to new investigations, generating further experience for subsequent learning. We also introduce DFM-Bench, which provides comprehensive assessment of hypothesis quality and cumulative research progress at the levels of hypotheses, revisions, development paths and research runs. Across four life-science domains, automated evaluation shows that hypothesis quality rises from 4.86 for the base model to 5.39 after two rounds of self-improvement, with the large gains in falsifiability and data resolvability. We further validate hypothesis quality through expert review, comparing our **DiscoveryRSI-27B** with other research agents. Preliminary human assessments show that DiscoveryRSI-27B scores higher than GPT-6 Luna on hypothesis importance and whether claims are plausible and evidence-bounded. DiscoveryRSI-27B achieved a 6.3% rate of hypothesis correspondence with research published after the model cutoff, largely outperform GPT-6 with 2.5% and Claude Opus 5 with 2.3%, while delivering greater human-rated improvement along research paths.

1. Introduction

Research requires developing worthwhile hypotheses into practical investigations. Developing such hypotheses involves gathering evidence, revising explanations, and deciding which directions merit further effort. This process generates experience about both the hypotheses under investigation and the decisions that guide their development, which allows human scientists to better design their next study. For research agents, using this experience to improve future research requires identifying useful contributions and translating them into changes to the model and research workflow. This motivates our central research question: *How can research agents turn hypothesis development and investigation into a recursive learning process that improves subsequent research?*

Addressing this question requires connecting hypothesis development and experimental execution with learning that improves the agent over investigations. Co-Scientist [7] develops hypotheses through multi-agent evolutionary refinement, while AI Scientist-v2 [32] connects hypothesis generation with experimental execution through agentic tree search. ScienceIDE [6] constructs learning environments from scientific codebases, and ScienceBuddy [31] turns researcher interactions into executable tasks and feedback for model and harness improvement. Building on these approaches, we focus on learning from hypotheses that agents actively develop: preserving how proposals evolve, link-

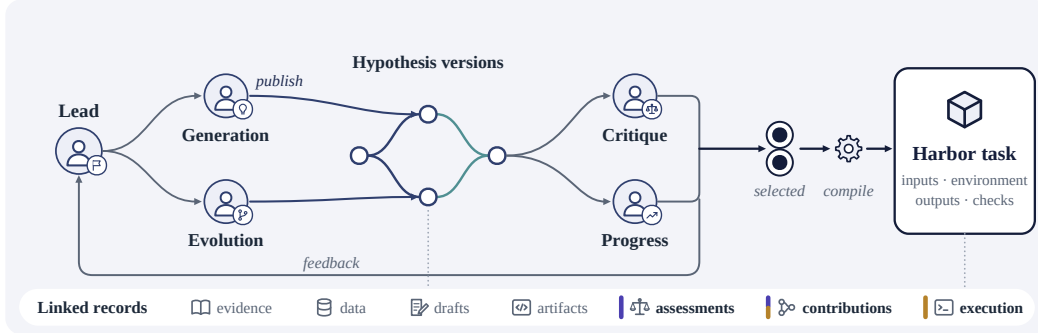
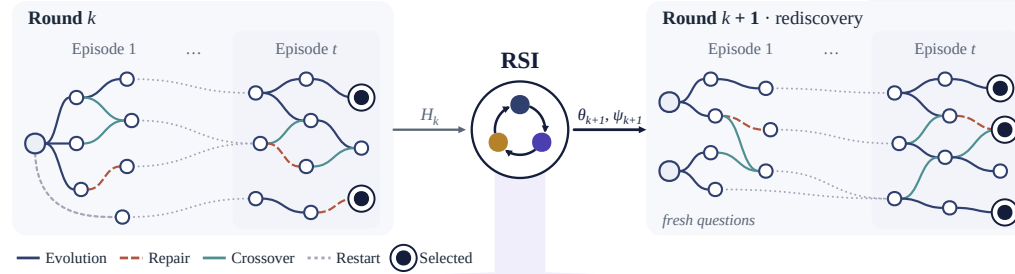
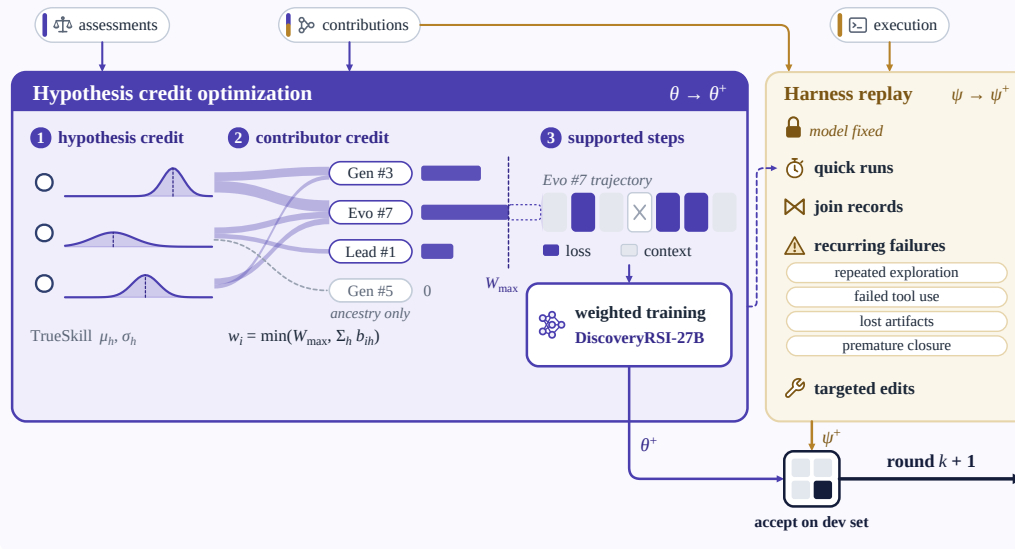
a Multi-agent search in episode t **b Research trees across RSI rounds****c Recursive self-improvement step**

Figure 2 | DiscoveryRSI couples multi-agent hypothesis development with recursive self-improvement. **(a)** Within an episode, a Lead coordinates Generation and Evolution; candidate hypothesis versions are critiqued and progress is recorded. Selected hypotheses are compiled into Harbor tasks, with linked evidence, data, drafts, artifacts, assessments, contributions, and execution records. **(b)** Research trees carry hypotheses across RSI rounds: the recorded state H_k enters RSI, which produces updated parameters $(\theta_{k+1}, \psi_{k+1})$ and fresh questions for rediscovery. **(c)** Two update branches learn from the same experience: *hypothesis credit optimization* turns assessments and documented contributions into bounded weighted training, while *harness replay* keeps the model fixed, diagnoses recurring failures, and tests targeted workflow edits. Accepted updates are checked on the development set before round $k + 1$.

ing assessments of research quality to the contributions behind them, and using this experience to improve subsequent investigations.

Turning hypothesis development into a recursive learning process presents challenges in both execution and learning. Translating a proposal into an executable investigation requires resolving its data requirements and verification criteria while preserving the research question. Learning from the resulting research experience involves a further challenge: a hypothesis’s final assessment does not specify which contributions should be reinforced. One agent may propose an explanation, another introduce a useful control, and a third revise the investigation plan in response to new evidence. Their contributions can remain obscured if only the final proposal is retained. Addressing these challenges involves making investigation plans concrete while preserving the development histories that connect assessments to the actions behind them.

To address these challenges, we introduce **DiscoveryRSI**, a general framework that turns hypothesis-driven research into a recursive learning process (Figure 2). It develops research trees of testable hypotheses through multi-agent evolutionary search, preserving the evidence and decisions behind each proposal and revision. Selected computational investigations are compiled into executable Harbor tasks with concrete inputs, environments, required outputs, and verification checks. The retained research histories support three coupled recursive improvement mechanisms. **Hypothesis credit optimization** links assessments of research quality to documented contributions and assigns bounded credit to the contributing histories, providing supervision for the associated model actions. We use this supervision to train **DiscoveryRSI-27B**. **Harness replay** examines these histories to identify productive strategies and recurring failures, informing changes to agent coordination, evidence access, and research allocation. **Rediscovery** applies accepted model and harness updates to fresh investigations, generating new hypotheses and experience for subsequent learning.

We introduce **DFM-Bench** to evaluate hypothesis quality and cumulative research progress. It assesses whether proposals are worthwhile and testable, what revisions contribute, and how development paths and research runs accumulate evidence and explore different directions. These assessments use hypotheses and their supporting research records, with the execution readiness of compiled Harbor tasks measured separately.

Our contributions are fourfold:

- **An operational definition and a general framework for hypothesis-driven research.** We define valuable and testable research hypotheses as joint objects comprising a claim, an investigation plan, decision criteria, and alternative explanations (Definition 1). DiscoveryRSI develops these objects through multi-agent evolutionary search, compiles selected investigations into executable Harbor tasks, and preserves the underlying research decisions and development histories as reusable learning experience.
- **Recursive self-improvement from research experience.** We develop hypothesis credit optimization to attribute assessments of research quality to documented contributions and train DiscoveryRSI-27B, harness replay to refine research strategies and workflows, and rediscovery to apply accepted improvements to subsequent investigations.

- **A benchmark for hypothesis quality and cumulative research progress.** DFM-Bench evaluates hypotheses, revisions, development paths, research runs, and research contributions. We further evaluate DiscoveryRSI through expert-supervised review and controlled model comparisons with its base model and other research agents. Results show clear gains over rounds of self-improvements and changes in agent behavior patterns.
- **Empirical evidence for learning from scientific research.** DiscoveryRSI increase mean hypothesis quality from 4.86 to 5.39, with gains in falsifiability, alternative explanations, and data resolvability. Expert-supervised assessment shows greater human-rated improvement along research paths, with an average gain per revision of +0.07, compared with +0.03 for GPT-5.6 Sol and −0.01 for GPT-6 Luna. DiscoveryRSI-27B also achieves a 6.3% rate of hypothesis correspondence with post-cutoff research, compared with 2.5% for GPT-6 and 2.3% for Claude Opus 5. Together, these findings show that learning from research histories improves the formulation of testable investigations and the development of hypotheses toward concrete research opportunities.

2. Valuable and Testable Research Hypotheses

2.1. An Operational Definition

We introduce a joint operational definition of **valuable and testable research hypotheses** for hypothesis-driven research agents. It makes the scientific value of a claim and the evidential capacity of its investigation explicit within one revisable research object.

Definition 1 (Valuable and testable research hypothesis) Relative to a documented evidence base $\mathcal{E}$ and declared data, experimental, and resource constraints $\mathcal{B}$, a research hypothesis is represented as

$$\mathcal{H} = (c, \pi, d, a), \quad (1)$$

where c is an evidence-grounded claim, π an investigation plan, d its decision criteria, and a the relevant alternative explanations. It is **valuable** when resolving the uncertainty in c could make a consequential addition to current understanding or change a justified research decision relative to $\mathcal{E}$. It is **testable** when π specifies an investigation feasible under $\mathcal{B}$ whose observable outcomes, interpreted through d and a , can provide evidence that supports, weakens, or bounds c , with unresolved distinctions stated explicitly. A valuable and testable research hypothesis satisfies both requirements.

The four components make this definition operational. The **claim** specifies the system or population, scope, observable quantities, and scientific uncertainty. The **investigation plan** identifies the data or measurements, comparisons, controls, analysis, and required dependencies. The **decision criteria** state, before interpreting results, which outcomes would support or weaken the claim and which would remain inconclusive. The **alternative explanations** identify competing accounts, confounding, or measurement limitations, together with the comparisons that can distinguish them and the limits of those comparisons.

Scientific value can arise from a new explanation, a useful replication, a measurement check, or an informative negative result when it changes understanding or research priorities. Testability depends on the declared evidence and constraints: unavailable measurements or experiments remain explicit dependencies until they can be obtained.

These requirements assess a proposed research opportunity; they do not establish that the claim is true. DFM-Bench evaluates the resulting objects through its scientific-quality profile (Section 4), while completed execution and independent support are recorded separately.

2.2. From Observations to Research Questions

DiscoveryRSI focuses on how a research question is formed and tied to available evidence. IdeaScientist develops proposals from a specified problem and challenge, including evaluation plans and falsifiable predictions [18]. Our emphasis is the joint development of a hypothesis and the investigation that can examine it: observations and data constraints shape the question, its comparisons, alternative explanations, and interpretation.

Descriptive analyses characterize observed patterns, variation, and measurement coverage, helping identify questions worth pursuing and the evidence available to investigate them. Each proposal then ties its claim to a defined system or population, observable quantities, and interpretable outcomes.

2.3. A Hypothesis with an Investigation Plan

Figure 3 instantiates Definition 1 with a MIMIC-IV proposal. Its value lies in asking whether repeated creatinine–urine output mismatch adds information about adverse outcomes after ICU discharge. Its testability comes from a defined population and outcome window, patient-level held-out comparisons against simpler baselines, explicit interpretation criteria, and alternative explanations that bound the claim. A supported lack of added information would also inform whether more complex models merit further investigation.

Recorded hypothesis versions retain links to their predecessors and supporting evidence. These records make changes in the question, investigation plan, and interpretation traceable, providing the research histories used by the trees and learning mechanisms in the following section.

3. DiscoveryRSI

DiscoveryRSI develops research hypotheses across episodes and retains the work behind each version as linked research records (Figure 2). These records connect scientific assessments to contributing actions for model training, and connect execution failures to changes in the research harness. Selected investigations become executable Harbor tasks. Hypothesis credit optimization, harness replay, and rediscovery then turn this accumulated experience into a recursive learning process. Algorithm 1 summarizes the complete protocol, from within-run hypothesis development to accepted updates and fresh collection.

3.1. Research Trees and Hypothesis Development

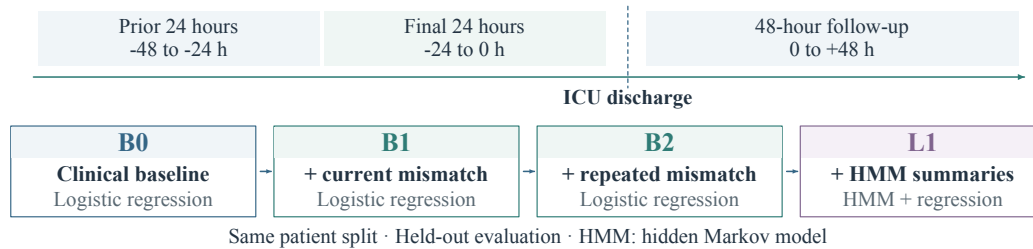
Research hypothesis trees. Each node is a versioned hypothesis with its investigation plan, decision criteria, alternative explanations, and evidence. Parent links record scientific development: a revision has a predecessor, while a combination can have several parents. The ancestry can therefore form a directed acyclic graph, although we

Do repeated kidney-recovery mismatches predict risk after ICU discharge?

Recorded MIMIC-IV proposal · Adult ICU survivors · 48-hour in-hospital outcomes

a Research hypothesis

Creatinine, a kidney blood marker, falls while urine output stays low. Repeated mismatch may identify higher risk of ICU readmission or in-hospital death within 48 hours after ICU discharge.

b Investigation plan**c Decision criteria**

- **Support:**
Stable risk information beyond the baseline.
- **Weaken:**
Signal fades after adjustment.
- **Unclear:**
Sparse events or unstable estimates.

d Alternative explanations

- Sicker patients may have both mismatch and worse outcomes.
- Treatments may change creatinine and urine output.
- Discharge decisions may select patients with different risks.
- Frequent testing may make mismatch easier to detect.

Figure 3 | A recorded research hypothesis and its planned investigation. This MIMIC-IV proposal concerns adults surviving a first ICU stay of at least 48 hours. **(a)** It asks whether repeated creatinine–urine output mismatch adds information about ICU readmission or in-hospital death within 48 hours after ICU discharge, during the same hospitalization. **(b)** B0–B2 compare clinical, current-mismatch, and repeated-mismatch logistic models; L1 adds hidden Markov model summaries, subject to data-support and stability checks. Models share a patient-level training, tuning, and test split. **(c,d)** Decision criteria and alternatives bound interpretation. No clinical outcome is asserted.

use research tree as an intuitive description. It spans episodes and remains distinct from the sequence of agent actions that produced each node.

Linked research records. The hypothesis tree is part of a broader network linking hypotheses to sources and data, drafts and supporting artifacts, scientific assessments, agent contributions, and execution records. Files preserve the research material; registry entries, identifiers, paths, and available content hashes preserve its relationships. A shared artifact can inform several hypotheses without making them scientific descendants of one another. These links let us trace an assessed contribution to the histories that produced it, or trace a recurring failure to the decisions and execution events behind it. Appendix A specifies the record types and link semantics.

Growing through evolutionary operations. Following operator-based search such as Frontis-MA1 [33], evolution develops a hypothesis, repair addresses an identified defect, and crossover combines contributions from multiple parents. Scientific restart opens a new question without assigning parents; earlier work may remain as context. Lead selects directions, parents, and operations. Generation and Evolution investigate evidence and publish proposals or revisions. Blinded Critique compares scientific merit, while ancestor-aware Progress assesses substantive progress, necessary repair, repetition,

or insufficient evidence. Their assessments guide subsequent work and are retained in the linked records.

3.2. From Hypotheses to Executable Investigations

Selected computational plans are compiled into Harbor tasks [9], with concrete inputs, environments, required outputs, and verification checks. Compilation preserves the hypothesis’s scientific commitments: the population or biological system, proposed comparison, discriminating outcomes, and interpretation criteria. Unavailable data and physical experiments remain explicit dependencies.

Execution checks establish whether a task satisfies its computational contract; scientific assessment determines what the investigation can establish. The compiled package remains linked to its selected hypothesis, inputs, artifacts, and validation records. Hypothesis development can therefore supply learning experience before a complete investigation has been executed.

3.3. Hypothesis Credit Optimization

Hypothesis credit optimization follows three links: scientific assessments determine hypothesis credit; contribution records allocate it to documented research histories; supported actions in those histories become training targets (Figure 2c).

Assessing hypotheses. Blinded Critique provides preferences, ties, or abstentions under compatible data and review conditions. TrueSkill [10] summarizes relative preference strength and uncertainty:

$$\begin{aligned} z_h &\sim \mathcal{N}(\mu_{h,0}, \sigma_{h,0}^2), & t_{h,e} \mid z_h &\sim \mathcal{N}(z_h, \beta^2), \\ P(a \succ b \mid z_a, z_b) &= \Phi\left(\frac{z_a - z_b - \varepsilon}{\sqrt{2}\beta}\right), \\ p(z_h \mid \mathcal{D}_{\text{cmp}}) &\approx \mathcal{N}(\mu_h, \sigma_h^2). \end{aligned} \quad (2)$$

Here z_h is rubric-conditioned preference strength, $t_{h,e}$ its noisy comparison performance, β the noise scale, ε the tie margin, and Φ the standard normal CDF. Insufficient evidence supplies no preference update. The resulting (μ_h, σ_h) describe comparative assessment, not the probability that a hypothesis is true. Separate Progress judgments assess changes relative to actual predecessors (Appendix B.1).

Allocating contributor credit. Assessment coverage, uncertainty, validity, and recorded progress determine a bounded hypothesis weight. The linked contribution records then identify histories that produced a proposal, supplied a supporting investigation, or made a documented coordination decision. One hypothesis can receive contributions from several Generation or Evolution histories and from Lead assignments. For history i , the accumulated credit is

$$w_i = \min\left(W_{\text{max}}, \sum_h b_{ih}\right), \quad (3)$$

where b_{ih} is the allocation from hypothesis h after contributor sharing, duplicate suppression, and scientific-family budget limits. A history is a particular recorded invocation, not all work by a role. Unsupported shares remain unallocated; ancestry or participation alone earns no credit. Appendices B and A.3 give the allocation rules and provenance requirements.

Reinforcing supported actions. Credit controls how strongly a history contributes to training. Within it, the target set selects the assistant responses and tool-call arguments supported by the recorded contribution. This can include several critical steps across different role histories. Tool observations and known failed actions remain available as context but receive no direct loss. We train DiscoveryRSI-27B with

$$\mathcal{L}(\theta) = - \sum_i \frac{\omega_i}{|U_i|} \sum_{u \in U_i} \log \pi_\theta(a_{iu} \mid x_{iu}), \quad (4)$$

where U_i contains eligible assistant-token positions, x_{iu} is their recorded context under the serialization policy in [Appendix C](#), and ω_i combines contributor credit with domain and shared-memory-group balancing. Thus, the history’s weight is distributed over its supported targets, rather than over every message. Histories without eligible targets are excluded. Allocation and serialization details appear in [Appendices B and C](#).

3.4. Harness Replay

Harness replay improves the research workflow while holding model weights fixed. Re-Act formalizes interleaved reasoning and acting [\[35\]](#); Self-Harness and Agentic Harness Engineering study improvement from deliberately minimal tool-using configurations [\[17, 37\]](#). Our agents start from mature Codex or Claude Code harnesses. Their session, context, filesystem, and tool-execution mechanisms provide the substrate for complex investigations that continue across many actions. We adapt research coordination, evidence access, review timing, and work allocation on top of that substrate ([Appendix A.4](#)).

Short diagnostic runs, including runs of a newly trained model candidate, first reveal where work breaks down. Agent contribution and decision records are joined with execution and delivery records to locate repeated exploration, failed tool use, lost artifacts, or premature closure. These diagnoses motivate targeted harness edits. For histories $\mathcal{H}_k$ collected with fixed model θ_k and harness ψ_k , candidate selection is

$$\begin{aligned} \Psi_k^{\text{cand}} &= \text{Propose}(\psi_k; \mathcal{H}_k), \\ \psi_k^+ &= \text{Select}_{\text{dev}}\left(\{\psi_k\} \cup \Psi_k^{\text{cand}} \mid \theta_k, B\right), \end{aligned} \quad (5)$$

where B denotes matched resource limits. Quick runs and supported replay diagnose problems and screen changes; matched development evaluation determines acceptance, considering scientific quality, exploration breadth, reliability, and cost. The incumbent remains eligible. Replay uses recorded observations and continuations; an unrecorded alternative requires execution from a compatible restored state. [Appendices A.4](#) distinguish diagnosis, replay, and controlled comparison.

3.5. Rediscovery

Rediscovery applies accepted model and harness updates to fresh investigations. Let θ_k^+ be the model candidate from hypothesis credit optimization and ψ_k^+ the harness candidate selected through replay and development evaluation. The next cycle is

$$\begin{aligned} (\theta_{k+1}, \psi_{k+1}) &= \text{Accept}_{\text{dev}}\left(\{\theta_k, \theta_k^+\} \times \{\psi_k, \psi_k^+\}\right), \\ \mathcal{H}_{k+1} &\sim \text{Research}\left(\theta_{k+1}, \psi_{k+1}; \mathcal{Q}_{k+1}^{\text{fresh}}, B\right). \end{aligned} \quad (6)$$

The acceptance procedure evaluates qualifying updates separately and jointly, so the model, harness, both, or neither may change. The accepted system returns to fresh

questions and collects new hypotheses, evidence, assessments, and contribution histories. Scientific restart changes direction within a research run; rediscovery changes the system used for the next run. Versioned records connect successive cycles while keeping development and final evaluation families separate (Appendices A.4).

Algorithm 1 DiscoveryRSI: recursive improvement from hypothesis development

Require: $(\theta_0, \psi_0), \{Q_k^{\text{fresh}}\}_{k=0}^{K-1}, Q_{\text{dev}}, Q_{\text{test}}, \text{budget } B$

- 1: **for** $k = 0, \dots, K - 1$ **do**
- 2: $(G_k, \mathcal{H}_k) \sim \text{Research}(\theta_k, \psi_k; Q_k^{\text{fresh}}, B)$ ▷ Tree and histories
- 3: $\mathcal{T}_k \leftarrow \text{CompileHarbor}(\text{Select}(G_k))$
- 4: $\{(\mu_h, \sigma_h, \Delta_h)\}_{h \in G_k} \leftarrow \text{Assess}(G_k, \mathcal{H}_k)$ ▷ Merit and progress
- 5: $\{b_{ih}\} \leftarrow \text{Credit}(\{(\mu_h, \sigma_h, \Delta_h)\}, \mathcal{H}_k)$ ▷ Bounded attribution; Appendix B
- 6: $w_i \leftarrow \min(W_{\text{max}}, \sum_h b_{ih})$
- 7: $U_i \leftarrow \text{SupportedTargets}(i, \mathcal{H}_k); \{\omega_i\} \leftarrow \text{Balance}(\{w_i : |U_i| > 0\})$
- 8: $\mathcal{L}_k(\theta) = - \sum_{i: |U_i| > 0} \frac{\omega_i}{|U_i|} \sum_{u \in U_i} \log \pi_\theta(a_{iu} \mid x_{iu})$ ▷ Eq. (4)
- 9: $\theta_k^+ \leftarrow \text{Update}(\theta_k; \mathcal{L}_k)$ ▷ ψ_k fixed
- 10: $\Psi_k^{\text{cand}} \leftarrow \text{Propose}(\psi_k; \mathcal{H}_k)$ ▷ Supported replay
- 11: $\psi_k^+ \leftarrow \text{Select}_{\text{dev}}(\{\psi_k\} \cup \Psi_k^{\text{cand}} \mid \theta_k, Q_{\text{dev}}, B)$ ▷ θ_k fixed
- 12: $(\theta_{k+1}, \psi_{k+1}) \leftarrow \text{Accept}_{\text{dev}}(\{\theta_k, \theta_k^+\} \times \{\psi_k, \psi_k^+\} \mid Q_{\text{dev}}, B)$
- 13: **end for**
- 14: $S_{\text{test}} \leftarrow \text{Eval}(\theta_K, \psi_K; Q_{\text{test}}, B)$ ▷ Untouched test families
- 15: **return** $(\theta_K, \psi_K), S_{\text{test}}$

4. DFM-Bench: Evaluating Scientific Quality and Research Progress

DFM-Bench evaluates scientific inquiry through five connected questions: whether a hypothesis is worthwhile and testable, what a revision contributes, whether a development path accumulates useful evidence, how a research run allocates effort, and what distinct scientific contribution emerges. Each question requires evidence appropriate to its level.

Table 1 | Evaluation coverage of selected scientific-discovery benchmarks. A check denotes explicit assessment, a circle related or partial assessment, and a cross no explicit assessment under our definitions. Recording intermediate actions alone does not establish coverage. Definitions and qualifications appear in Appendix D.6.

Benchmark	Scientific assessment		Research-process assessment			
	Hypothesis merit	Contribution maturity	Evidence use	Revision	Development path	Whole-run allocation
AgentIdeaBench [21]	✓	×	○	×	×	×
DiscoveryBench [20]	○	×	○	×	×	×
BLADE [8]	×	×	○	×	×	×
ScienceAgentBench [3]	×	×	○	×	×	×
OEB [27]	×	×	✓	✓	○	○
DFM-Bench (ours)	✓	✓	✓	✓	✓	✓

4.1. Five Levels of Scientific Assessment

Hypothesis quality. Is an individual hypothesis worthwhile and testable? Assessment examines the proposal, investigation plan, and supporting evidence. It produces a

scientific-quality profile covering importance, novelty, grounding, falsifiability, data resolvability, feasibility, alternative explanations, and actionability, together with an aggregate score when the required assessments are complete.

Revision value. What did a revision contribute? Assessment compares a hypothesis with its actual predecessor and examines the evidence and actions supporting the change. It distinguishes substantive progress, necessary repair, repetition, and insufficient evidence. Changes in hypothesis scores are recorded separately from these judgments.

Development-path progress. Does a development path accumulate useful evidence? Assessment follows an ordered sequence of actual ancestry relationships, examining evidence acquisition, coherent revision, scientific continuity, and decisions to continue or stop. Agent action histories provide supporting evidence for these judgments, while the development path records how the hypotheses themselves changed.

Research-run quality. Does a research run explore and allocate effort productively? Assessment examines the full research graph, selection and stopping decisions, and available opportunity and resource records. It considers the breadth and depth of exploration, synthesis across branches, and allocation of the research budget. Run-level quality is assessed directly rather than obtained by averaging path scores.

Scientific contribution. What distinct scientific contribution emerges across related proposals? Assessment groups proposals by scientific question and records what was added beyond the starting evidence: a new distinction, an informative negative result, a supported finding, or a reduction in uncertainty. Contribution maturity distinguishes proposed ideas from contributions supported by completed investigations and independent checks.

4.2. Coverage Relative to Existing Benchmarks

Existing benchmarks examine complementary aspects of scientific work, including ideation, goal-directed discovery, analytical decisions, and executable tasks [3, 8, 20, 21]. OEB further evaluates evidence use, experimentation, and research-process integrity [27]. DFM-Bench makes hypotheses, revisions, development paths, research runs, and contribution families explicit assessment objects. Table 1 compares their evaluation coverage; the accompanying definitions and qualifications appear in Appendix D.6.

5. Expert-Supervised Assessment of Scientific Hypotheses

We conduct expert-supervised human assessment to examine the scientific value and design quality of generated hypotheses. The review considers whether a proposed investigation addresses a worthwhile question, fits the available data, distinguishes alternative explanations, and yields interpretable outcomes.

The assessment involved 27 medical doctors, medical students and bioscience PhD students with prior biomedical or clinical research backgrounds, under three senior supervisors. The review cohort contains 292 proposals from six dataset groups spanning the four life-science domains: HCC clinical cohort (52), MIMIC-IV (60), eICU (60), UK Biobank (55), GEO targets (35), and GEO mechanisms (30). Proposals were drawn from a fixed hypothesis-atlas snapshot and presented in randomized order.

Anonymous proposal packets. Reviewers received English proposal copies containing the scientific question, investigation plan, data requirements, alternative explanations, and supporting citations. Producer and execution provenance were withheld, while scientific methods and references were retained. AI reviewer grades were hidden, and reviewers were instructed to assess each proposal on its own merits without consulting its revision history in the atlas. The interface presented the proposal alongside the assessment form (Figure 4); Appendix E consolidates the review protocol.

Scientific judgments and design flaws. The protocol asks reviewers to record an overall research-value judgment first, using a five-point scale from not worth pursuing to top priority, together with a recommendation to prioritize, revise, defer, or reject the proposal. Reviewers then assess six criteria on anchored four-point scales: importance, advance over known work, plausible and evidence-bounded claims, data fit, alternative explanations, and informative results. A separate verdict identifies whether a consequential design flaw is absent, fixable, or fatal. “Can’t judge” is available when evidence or relevant expertise is insufficient. A flaw checklist, a description of the main flaw, and further comments are optional.

a Anonymous scientific proposal

Unresolved clinical question and hypothesis

After an ICU patient develops creatinine-defined AKI, clinicians may infer recovery when serum creatinine falls. The unresolved question is whether that inference is unsafe when urine output and recorded fluid balance do not improve at the same time.

Hypothesis: among adults with a first eligible ICU stay who develop creatinine-defined AKI, survive and remain in the ICU to a fixed 48-hour post-onset landmark, and show a prespecified creatinine fall, patients with low primary-set urine output or positive recorded-net-volume proxy at that landmark have a higher 7-day risk of hospital death or ICU-recorded incident kidney-replacement therapy than patients with both urine recovery and non-positive fluid balance. The primary estimand is the standardized prognostic risk difference, discordant minus concordant recovery.

This tests a falsifiable association. It does not claim that discordance causes death, that positive balance proves hemodilution or congestion, or that the patient has measured persistent renal dysfunction. The clinically consequential decision it could inform is whether a falling creatinine alone should be sufficient for de-escalating renal monitoring or transition planning after AKI. The result would be hypothesis-generating for a monitoring/intervention study, not a bedside rule.

b Structured assessment

1 Overall judgment

Research value (1–5) and recommendation

- ☒ 5 **Top priority** — Among the best questions this dataset could address; I would want this done.
- ☐ **Prioritize** — Pursue now.

2 Six anchored criteria

Each criterion: 1–4 or “Can’t judge”

- Importance
- Advance over known work
- Plausible, evidence-bounded claims
- Data fit
- Alternative explanations
- Informative result

3 Design-flaw verdict

None · fixable · fatal · “Can’t judge”

- ☒ **No** — No flaw that would change what the main result means.

Optional: flaw checklist and comments

Figure 4 | Expert-supervised assessment of scientific hypotheses. (a) An anonymous MIMIC-IV proposal presents its scientific question, planned comparison, and claim limits. (b) Reviewers record an overall judgment, assess six anchored criteria, and identify design flaws; flaw details and comments are optional. The criterion list summarizes the rubric, while radio-option excerpts reproduce one saved assessment. Identifying metadata is omitted. The example illustrates the review procedure rather than an aggregate evaluation result.

6. Experimental Setup

Research domains. Our main evaluation of DiscoveryRSI covers four life-science domains: clinical and population health, therapeutic target prioritization, disease mechanisms, and population multi-omics. Inputs include the HCC clinical cohort, MIMIC-IV [12], eICU [24], UK Biobank [30], and molecular studies from GEO and ArrayExpress [5, 26]. To

test generalization beyond life science, we add a battery setting with four tasks: electrical diagnostics, health prediction, degradation mechanisms, and usage history and aging, specified in [Figure 5](#). Agents inspect data and literature, formulate hypotheses, and develop investigation plans specifying comparisons, measurements, and interpretation criteria.

Hypothesis-credit training. We perform full-parameter fine-tuning of DiscoveryRSI-27B with a 256K-token context window using hypothesis-credit-weighted supervision, using Qwen3.8-27B as initialization. From 15,792 collected research trajectories, attribution, filtering, and tokenization yield 2,397 training trajectories: 634 Lead, 861 Generation, and 902 Evolution histories. The training sequences contain 291.12 million tokens, of which 6.18 million (2.12%) receive direct supervision. We train for one epoch using Adam with a learning rate of 5×10^{-6} , a global batch size of 16, and BF16 precision. Detailed trajectory accounting, credit allocation, and token masking appear in [Appendix C](#).

Evaluation. We examine four questions: whether hypothesis quality improves across research rounds, how research behavior changes, whether human reviewers recognize useful revisions, and whether hypotheses inspire subsequent research. Automated quality evaluation covers 241 research questions across the four domains using GPT-5.6 Sol as the judge. Behavioral analysis examines six campaigns of 20 episodes per evaluated model variant. Human assessment covers 292 proposals reviewed by 27 participants with biomedical or clinical research backgrounds under three senior supervisors. Proposals are presented in randomized order with model identities and automated scores hidden ([Section 5](#)). Literature-based evaluation uses a shared reference corpus with model-specific temporal cutoffs.

7. Results

7.1. RQ1: Does learning from research histories improve hypothesis quality?

Hypothesis quality rises with each round of self-improvement. Mean automated hypothesis quality increases from 4.86 for the base model to 5.18 after the first research round and 5.39 after the second ([Figure 6](#)). The second-round model exceeds the base model in all four domains, with gains ranging from 0.28 points in disease mechanisms to 0.94 points in population multi-omics. The additional gain decreases from 0.32 in the first round to 0.21 in the second.

Improvements concentrate on the specification of testable investigations. Falsifiability rises from 4.00 to 6.20, consideration of alternative explanations from 4.44 to 5.98, data resolvability from 3.73 to 4.99, and feasibility from 4.79 to 5.87. These changes indicate more explicit tests, better-defined evidence requirements, and clearer interpretation of possible outcomes. Gains are uneven: importance declines from 6.48 to 6.04 and novelty from 4.87 to 4.58. The observed improvement therefore concerns investigation design more strongly than the perceived significance or originality of research questions.

7.2. RQ2: How does DiscoveryRSI change research behavior?

DiscoveryRSI allocates more effort to developing existing hypotheses. Evolve and repair operations increase from 77.5% of episodes for the base model to 90.0% after the first

A shared research process across life-science & battery domains

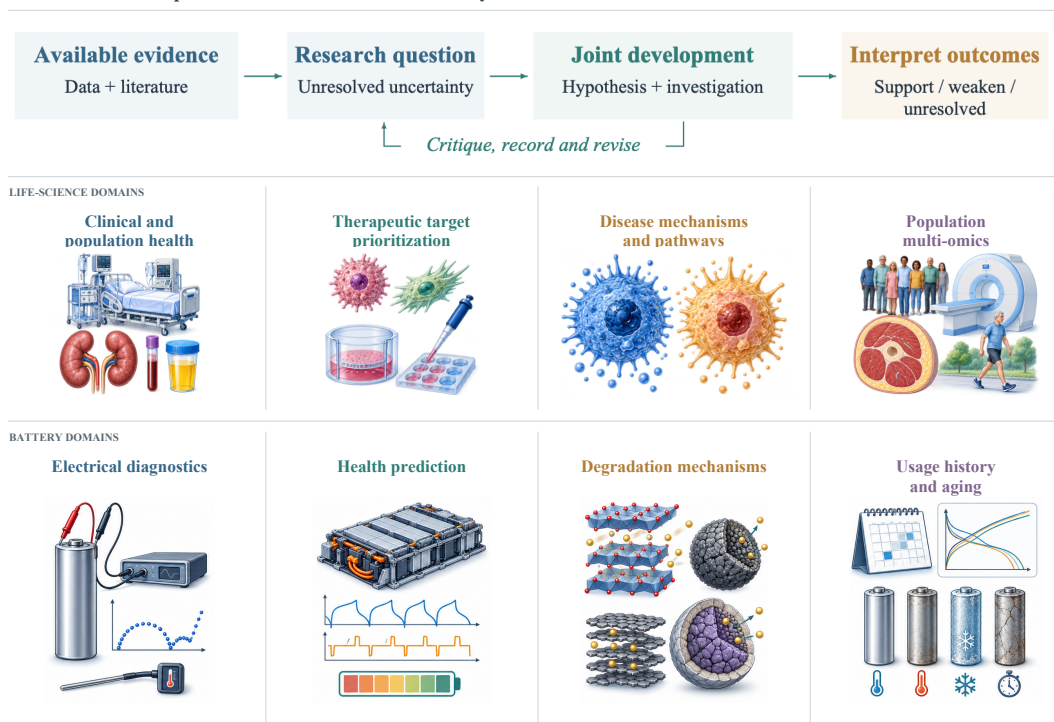


Figure 5 | Shared research process across life-science and battery domains. The process strip summarizes how evidence becomes a question, investigation, and evidence-bounded interpretation with iterative critique. The two rows show the four life-science domains and four battery domains used in the experiments.

round and 91.7% after the second. Restarts decrease from 21.7% to 5.8% and 5.0%, respectively (Figure 7). The shift is especially pronounced in episodes 11–20, where restarts fall from approximately 32% for the base model to 3% after the second round.

Hypothesis development paths become more evidence-driven. Development-path scores increase most for evidential and inferential discipline (+1.25), followed by technical and knowledge exploration (+0.34) and scientific continuity and synthesis (+0.24). Changes in cumulative progress, stopping decisions, and budget efficiency remain below 0.1 points, while disciplinary connections decline by 0.28. Together, these results suggest more sustained refinement and stronger evidential discipline, with limited improvement in broader exploration or resource allocation. Operation frequencies describe the research strategy; their value depends on the quality of the resulting revisions.

7.3. RQ3: Do human reviewers recognize improvements in the hypotheses?

Human assessment identifies a favorable balance of revision outcomes for DiscoveryRSI-27B. Within the assessed research trees, 41% of revisions improve the mean human score, 47% leave it unchanged, and 12% reduce it. The corresponding proportions are 31%, 37%, and 32% for GPT-5.6 Sol, and 25%, 44%, and 31% for GPT-6 Luna. Mean score change per revision is +0.07 for DiscoveryRSI-27B, compared with +0.03 and −0.01 for the two comparison models (Figure 8). These results measure progress within research trees and provide complementary evidence to the comparisons across research rounds.

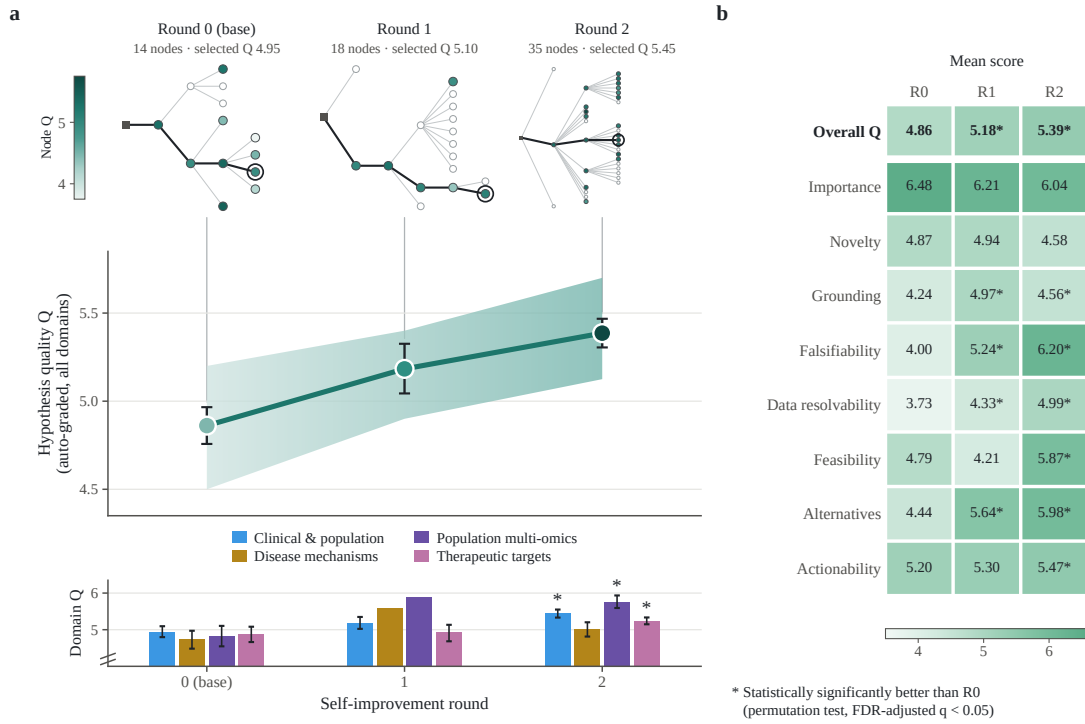


Figure 6 | Automated hypothesis quality across self-improvement rounds. (a) Top: example research tree from each round; color shows weighted quality across eight evaluation axes. Middle: mean hypothesis quality per round. Error bars show 95% confidence intervals and shading shows the interquartile range. Bottom: mean quality per domain with 95% confidence intervals. **(b)** Mean and axis-specific scores. All scores are produced by the automated grader using the weights in [Appendix D](#).

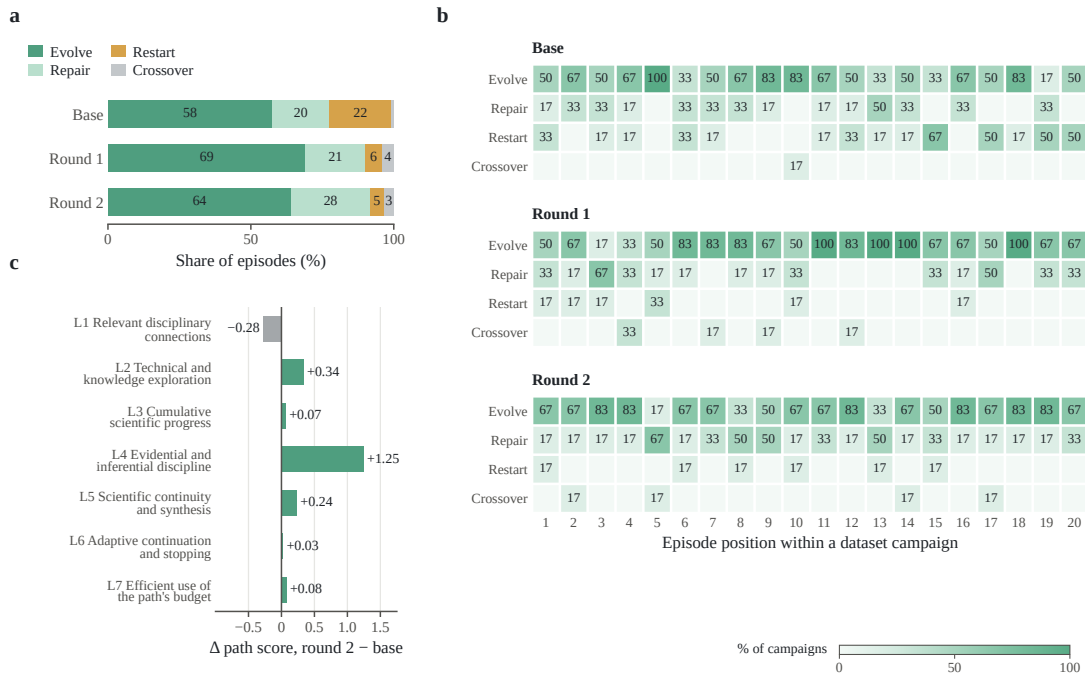


Figure 7 | Changes in research behavior across rounds. (a) Share of each action across all episodes in a round. **(b)** Share of campaigns taking each action at each episode position, sampled from six campaigns per round; blank cells indicate 0%. **(c)** Change in development-path scores from the base model to round 2.

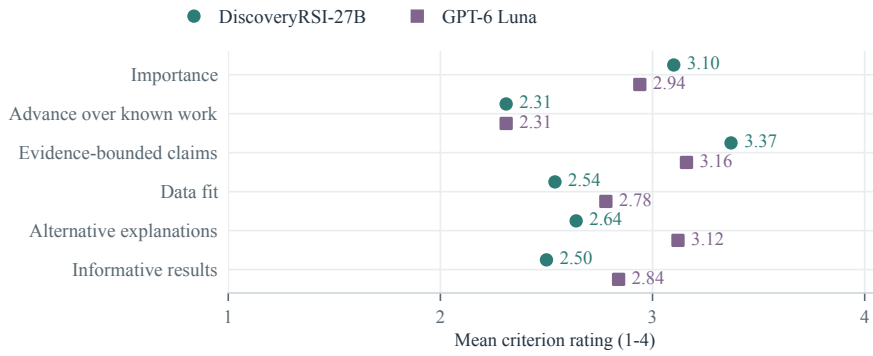
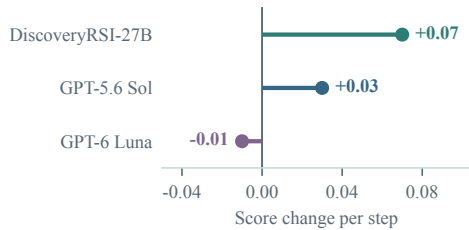
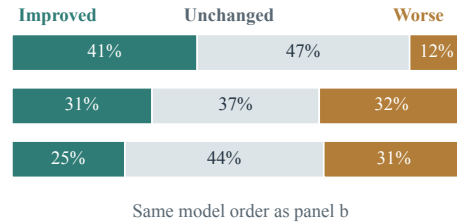
a Human assessment of hypothesis quality**b Mean revision change****c Revision outcomes**

Figure 8 | Human assessment of hypothesis quality and revision progress. (a) Mean ratings across six criteria on a 1–4 scale. **(b)** Mean score change per revision. **(c)** Fractions of improved, unchanged, and worse revisions, in the same model order as panel (b).

Absolute quality ratings show a mixed profile. DiscoveryRSI-27B scores higher than GPT-6 Luna on importance (3.10 versus 2.94) and plausible, evidence-bounded claims (3.37 versus 3.16). Both receive 2.31 for advance over known work, while GPT-6 Luna scores higher on data fit, alternative explanations, and informative results. Thus, the stronger revision outcomes coexist with remaining weaknesses in individual proposals.

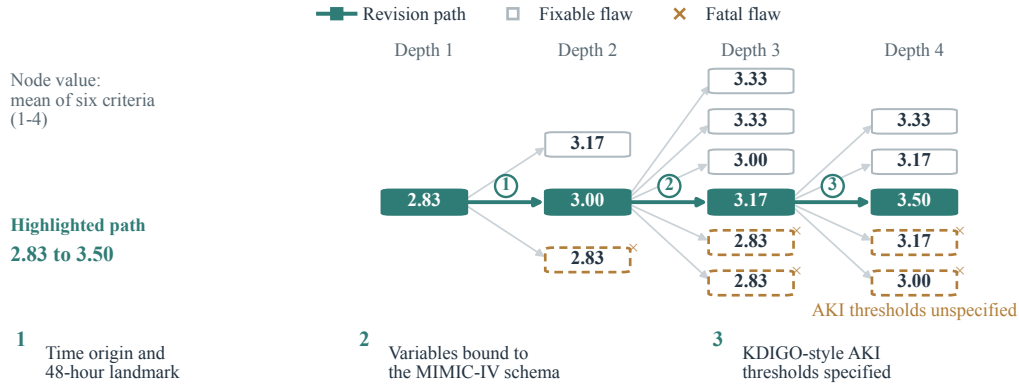
A revision path in MIMIC-IV. A MIMIC-IV development path illustrates how useful revisions improve an investigation (Figure 9). Successive versions clarify the time origin and 48-hour landmark, bind variables to available data fields, and specify acute kidney injury thresholds. The mean of six human-rated criteria increases from 2.83 to 3.50; data fit and informative results each improve by two points. This example connects revision gains to concrete design changes. Other branches retain consequential flaws, including two depth-four siblings with unspecified acute kidney injury thresholds, underscoring the need to evaluate each resulting proposal.

7.4. RQ4: Do the hypotheses inspire subsequent research?

We define “inspiration” as a concrete connection between a generated hypothesis and subsequent literature. A publication after the recorded cutoff must provide an experimental design, comparator, measurement, or testable evidence gap relevant to developing or testing the hypothesis.

Across the two pooled DiscoveryRSI production runs, 74 of 1,176 hypothesis versions meet this criterion, giving an inspiration rate of 6.3%. The corresponding rates are 5.4% for DeepSeek (14/257), 2.8% for GLM (1/36), 2.5% for GPT (59/2,341), and 2.3% for Claude (3/131). DiscoveryRSI reaches 16.9% in therapeutic targets, 7.4% in population multi-omics, 6.8% in disease mechanisms, and 2.3% in clinical and population health (Figure 10).

a MIMIC-IV revision path and assessed siblings



b Criterion ratings along the highlighted path

Importance	4	4	3	3
Advance over known work	3	3	3	3
Evidence-bounded claims	4	4	4	4
Data fit	1	2	2	3
Alternative explanations	3	3	3	4
Informative results	2	2	4	4
	Depth 1	Depth 2	Depth 3	Depth 4

Rating scale: 1 lowest, 4 highest

Figure 9 | Design changes and human ratings along a revision path. (a) A MIMIC-IV path and assessed siblings; node values are six-criterion means on a 1–4 scale. Numbered changes are explained below the tree. Dashed ochre nodes mark fatal flaws; the two final fatal siblings lack specified AKI thresholds. (b) Criterion ratings along the highlighted path.

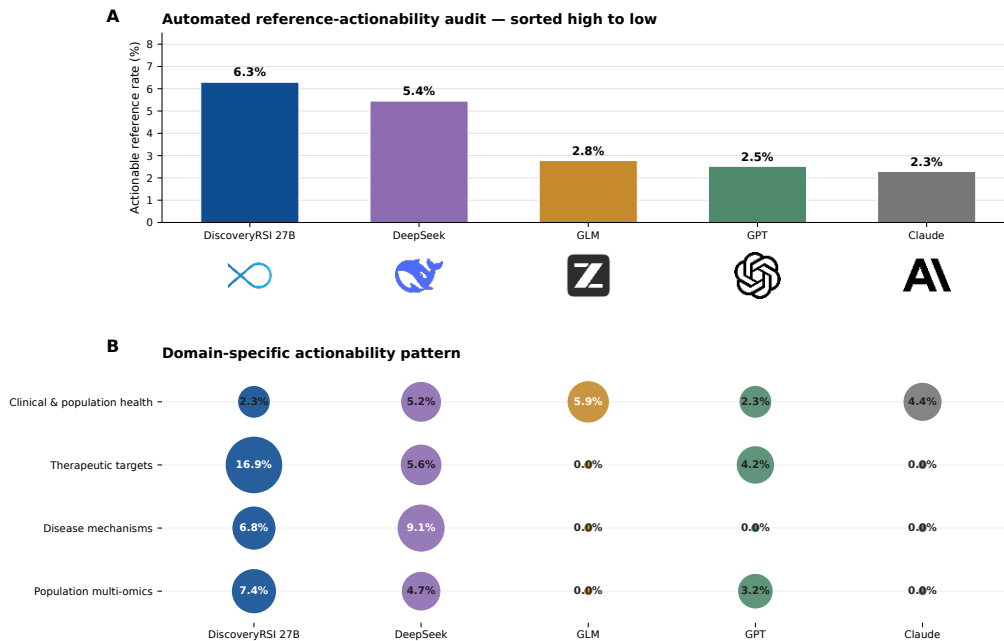


Figure 10 | Hypothesis inspiration measured through subsequent literature. The upper panel orders model groups by the proportion of generated hypothesis versions meeting the inspiration criterion. The lower panel reports the same rate across four domains. DiscoveryRSI-27B pools two production runs; Qwen original is omitted. The figure's reference-actionability labels denote this retrospective inspiration measure.

These descriptive results indicate that DiscoveryRSI hypotheses more frequently connect to concrete research directions in the retrieved literature. Group sizes, domain composition, temporal cutoffs, and retrieval coverage differ, and multiple versions may belong to the same research family. Prospective tracking is needed to determine whether researchers adopt these hypotheses and pursue the proposed investigations.

Literature comparison protocol. We construct a shared reference corpus through 420 backward and forward citation traversals, yielding 17,790 citation edges and 16,440 resolved publications. References are filtered using each model group’s recorded cutoff. The inspiration rate is the proportion of generated hypothesis versions receiving a positive automated judgment under the above definition:

$$R_{\text{insp}} = \frac{\sum_h a_h}{|\mathcal{H}|}, \quad (7)$$

where a_h indicates a positive judgment and $\mathcal{H}$ contains all generated hypothesis versions in the group. Corpus construction, cutoff handling, retrieval, and auditing are detailed in [Appendix F](#).

Why does Claude Opus 5 have a lower inspiration rate? We attribute part of Claude Opus 5’s lower measured rate to retrieval coverage and proposal format. Only 15 of its 131 proposals pass the retrieval gate (11.5%), compared with 355 of 1,176 for DiscoveryRSI-27B (30.2%). Among these retrieved proposals, however, the positive-judgment fractions are similar: 20.0% and 20.8%, respectively. Claude’s median proposal length is 41,134 characters, while retrieval uses the first 6,000 characters and the semantic audit sees the first 2,500. Provenance and revision notes can therefore displace scientific content from the evaluated excerpt. In a diagnostic using proposal titles as queries while retaining the corpus, scoring method, date rule, and overlap threshold, Claude’s retrieval passage increases from 15 to 95 proposals. This sensitivity supports a format-related retrieval explanation, but does not establish additional matches meeting the inspiration criterion.

7.5. RQ5: How well does DiscoveryRSI transfer beyond life sciences?

To test transfer beyond life science, we perform an additional battery domain evaluation covering electrical diagnostics, health prediction, degradation mechanisms, and usage history and aging. The evidence pack covers capacity and impedance aging measurements [19], impedance-based health prediction [13, 41], degradation-mode analysis [14, 16], and path-dependent aging experiments [25]. Investigations separate measured quantities from fitted degradation labels and specify cell, condition, and protocol splits. The same Lead–Generation–Evolution–Critique–Compiler loop is used, with battery-specific checks for operating conditions, sensing, physical plausibility, and competing explanations.

The battery domain applies the five-level DFM-Bench evaluation to 200 episodes and 351 generated hypotheses, with three AI reviews per hypothesis. The current panel reports a mean hypothesis quality of 5.61 on 165 complete scores, substantive progress on 214 of 366 transitions (58.47%), a mean complete-trajectory score of 81.52 across 152 paths, and a research round score of 71.25. 20 contribution families were identified, with 7 carrying diagnostic or research evidence.

8. Related Work

Discovery agents and evaluation. Discovery agents increasingly support hypothesis generation, experimental planning, and computational investigation. Co-Scientist develops hypotheses through multi-agent refinement [7], AI Scientist-v2 connects research ideation with agentic experimental search [32], and Biomni integrates biomedical tools and data [11]. Discovery Foundation Models frame problem formation and evidence-grounded revision as learnable capabilities [34], while ScienceIDE constructs environments for learning from scientific workflows [6]. Evaluation spans scientific ideation in AgentIdeaBench [21], data-driven discovery in DiscoveryBench and BLADE [8, 20], executable investigations in ScienceAgentBench [3], and research-process assessment in OEB [27]. DiscoveryRSI connects hypothesis development to subsequent model learning, while DFM-Bench makes hypotheses, revisions, development paths, research runs, and contribution families explicit evaluation objects.

Learning from evaluated research experience. Reward-guided training provides a foundation for converting generated experience into model updates. RAFT selects highly rated samples for fine-tuning [4], ReST-EM alternates generation and feedback-based filtering [29], and advantage-weighted regression learns through weighted likelihood objectives [23]. SimpleTES further studies learning from scientific exploration trajectories [36]. These approaches motivate the use of assessed research histories as supervision. DiscoveryRSI addresses the attribution problem within multi-agent hypothesis development: an assessment concerns a scientific proposal, whereas learning targets belong to the histories that contributed to it. Hypothesis credit optimization links these levels through documented contributions, bounded credit allocation, and supervision restricted to supported assistant actions.

Persistent and recursive agent improvement. Reflexion, GEPA, and ACE retain experience through verbal memory, prompt evolution, and contextual playbooks [1, 28, 40]; harness optimization extends adaptation to agent workflows and implementations [17, 37]. Darwin Gödel Machine and Hyperagents study persistent program-level improvement [38, 39], while MASS couples multi-agent workflow optimization with learning from selected executions [15]. ScienceBuddy connects harness refinement with model learning [31], and RSI-Anything uses predictive world models to guide interaction and retained system updates [?]. DiscoveryRSI focuses on how experience accumulated while developing scientific questions can improve later investigations.

9. Conclusion

DiscoveryRSI connects hypothesis development with learning from documented research contributions. The present work reaches concrete hypotheses and compiled investigation plans. The next step is to determine which proposals merit investigation, using solvers and frameworks such as RSI-Anything [?] to execute, assess, and refine them. Closing the discovery loop further requires wet-lab validation and methods that turn supportive, negative, and inconclusive outcomes into improved research decisions. Extending hypothesis discovery and DFM-Bench to additional scientific domains will help establish how these capabilities generalize. Together, these directions advance the DFM vision [34] of agents that formulate worthwhile questions, investigate them, and learn from the resulting evidence.

References

- [1] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziem, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: reflective prompt evolution can outperform reinforcement learning. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.
- [2] Anthropic. How Claude Code Works, 2026. URL <https://code.claude.com/docs/en/how-claude-code-works>. Official documentation, accessed 2026-10-14.
- [3] Zirui Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=6z4YKr0GK6>.
- [4] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: reward ranked finetuning for generative foundation model alignment. *Trans. Mach. Learn. Res.*, 2023, 2023. URL <https://openreview.net/forum?id=m7p507zblY>.
- [5] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, 2002. doi: 10.1093/NAR/30.1.207. URL <https://doi.org/10.1093/nar/30.1.207>.
- [6] Hejia Geng, Zesen Huang, Haoyang Li, Wenbin Li, Koutian Wu, Zihan Zhou, Yuanbo Pang, Weihao Liu, Zigong Xu, Zhiping Li, Zongzheng Zhang, Chuanfei Dong, Jiankai Sun, Tianzhe Zheng, Fengyu Xie, Yue Ma, Yueheng Shi, Tong Xie, Zonglin Di, Xianrong Liu, Qucheng Gao, Yimin Liu, Jiaming Pan, Sheng Huang, Xiao-Han Ma, Lanqing Yuan, Zhenlin Zhu, Ziang Liu, Ziyang Xu, Junkai Wang, Kangkai Liang, Jiayi Xian, Zehong Zhao, Liuwei Xu, Jingxu Xie, Peijin Zhang, Qiang Gao, Chengyi Xing, Zhe Zhao, Xi Wang, Yaopeng Xing, Xing Meng, Zhenfei Yin, Yingcheng Wu, and Ling Yang. Scienceide: Turning world’s scientific codebase into agent learnable environments. *CoRR*, abs/2609.19134, 2026. doi: 10.48550/ARXIV.2609.19134. URL <https://doi.org/10.48550/arXiv.2609.19134>.
- [7] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with co-scientist. *arXiv:2502.18864v2*, 2026. URL <https://arxiv.org/abs/2502.18864v2>.
- [8] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: benchmarking language model agents for data-driven science. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 13936–13971. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.815. URL <https://doi.org/10.18653/v1/2024.findings-emnlp.815>.
- [9] Harbor contributors. Harbor: Task format. Online documentation. URL <https://docs.harborframework.com/tasks/overview>. Accessed 2026-10-09.

- [10] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskilltm: A bayesian skill rating system. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 569–576. MIT Press, 2006. URL <https://proceedings.neurips.cc/paper/2006/hash/f44ee263952e65b3610b8ba51229d1f9-Abstract.html>.
- [11] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.
- [12] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [13] Penelope K. Jones, Ulrich Stimming, and Alpha A. Lee. Impedance-based forecasting of lithium-ion battery performance amid uneven usage. *Nature Communications*, 13(1):4806, August 2022. doi: 10.1038/s41467-022-32422-w. URL <https://doi.org/10.1038/s41467-022-32422-w>.
- [14] Niall Kirkaldy, Mohammad A. Samieian, Gregory J. Offer, Monica Marinescu, and Yatish Patel. Lithium-ion battery degradation: Comprehensive cycle ageing data and analysis for commercial 21700 cells. *Journal of Power Sources*, 603:234185, May 2024. doi: 10.1016/j.jpowsour.2024.234185. URL <https://doi.org/10.1016/j.jpowsour.2024.234185>.
- [15] Hyunin Lee, Jinglue Xu, Jeffrey Seely, Donghyun Lee, Somayeh Sojoudi, Matei Zaharia, and Yujin Tang. Recursive self-improvement through multi-agent self-supervision. *arXiv preprint arXiv:2610.12176*, 2026.
- [16] Ruihe Li, Niall D. Kirkaldy, Fabian F. Oehler, Monica Marinescu, Gregory J. Offer, and Simon E. J. O’Kane. The importance of degradation mode analysis in parameterising lifetime prediction models of lithium-ion battery degradation. *Nature Communications*, 16(1):2776, March 2025. doi: 10.1038/s41467-025-57968-3. URL <https://doi.org/10.1038/s41467-025-57968-3>.
- [17] Jiahang Lin, Shichun Liu, Chengjun Pan, Lizhi Lin, Shihan Dou, Xuanjing Huang, Hang Yan, Zhenhua Han, and Tao Gui. Agentic harness engineering: Observability-driven automatic evolution of coding-agent harnesses. *CoRR*, abs/2604.25850, 2026. doi: 10.48550/ARXIV.2604.25850. URL <https://doi.org/10.48550/arXiv.2604.25850>.
- [18] Jiarui Liu, Renjie Tao, Yiwei Liao, Chuanyang Jin, Kai Sun, Xiao Yang, Xinyuan Zhang, Xilun Chen, Zhuangqun Huang, Lechen Zhang, Yongjin Yang, Yinghui He, Weihao Xuan, Rakesh Wanga, Anuj Kumar, Mona T. Diab, Wen-tau Yih, and Xin Luna Dong. IdeaScientist: Orchestrating Agents for Grounded Scientific Ideation. *arXiv:2610.04074*, 2026. URL <https://arxiv.org/abs/2610.04074>.
- [19] Matthias Luh and Thomas Blank. Comprehensive battery aging dataset: capacity and impedance fade measurements of a lithium-ion NMC/C-SiO cell. *Scientific Data*, 11(1):1004, September 2024. doi: 10.1038/s41597-024-03831-x. URL <https://doi.org/10.1038/s41597-024-03831-x>.
- [20] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=vyflgpwfJW>.
- [21] Yunxiang Mo, Tianshi Zheng, Yisen Gao, Rui Wang, Newt Hue-Nam K. Nguyen, Kelvin Kiu-Wai Tam, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. Agentideabench: Benchmarking scientific ideation in the agent era. *CoRR*, abs/2609.07611, 2026. doi: 10.48550/ARXIV.2609.07611. URL <https://doi.org/10.48550/arXiv.2609.07611>.
- [22] OpenAI. Codex as a Platform: Build on the Open Agent Harness, 2026. URL <https://developers.openai.com/blog/codex-as-a-platform>. Official documentation, accessed 2026-10-14.
- [23] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019. doi: 10.48550/ARXIV.1910.00177. URL <https://doi.org/10.48550/arXiv.1910.00177>.
- [24] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care

- research. *Scientific data*, 5(1):180178, 2018.
- [25] Trishna Raj, Andrew A. Wang, Charles W. Monroe, and David A. Howey. Investigation of path-dependent degradation in lithium-ion batteries. *Batteries & Supercaps*, 3(12):1377–1385, September 2020. doi: 10.1002/batt.202000160. URL <https://doi.org/10.1002/batt.202000160>.
 - [26] Ugis Sarkans, Anja Füllgrabe, Ahmed Ali, Awais Athar, Ehsan Behrangi, Nestor Diaz, Silvie Fexova, Nancy George, Haider Iqbal, Sandeep Kurri, Jhoan Munoz, Juan Camillo Rada, Irene Papatheodorou, and Alvis Brazma. From arrayexpress to biostudies. *Nucleic Acids Res.*, 49(Database-Issue):D1502–D1506, 2021. doi: 10.1093/NAR/GKAA1062. URL <https://doi.org/10.1093/nar/gkaa1062>.
 - [27] Chengyang Shi, Xianglin Ji, Jintao Huang, Jicheng Wang, Yifeng He, and Jiachen Liu. Open-endedness bench: Measuring epistemic process from agent records. *arXiv preprint arXiv:2610.02588*, 2026.
 - [28] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.
 - [29] Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T. Parisi, Abhishek Kumar, Alexander A. Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL <https://openreview.net/forum?id=lNAYUngGFK>.
 - [30] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, 2015. doi: 10.1371/journal.pmed.1001779. URL <https://doi.org/10.1371/journal.pmed.1001779>.
 - [31] Shuhan Xue, Jianyuan Zhong, Ziyuan Nan, Wenbin Li, Zhaochen Yu, Jinchao Ding, Qiang Gao, Pengyu Zhan, Yuntong Zhang, Tian Cheng, Zhenfei Yin, Yingcheng Wu, and Ling Yang. Sciencebuddy: Recursive-in-recursive self-improvement for interactive scientific agents. *CoRR*, abs/2609.17523, 2026. doi: 10.48550/ARXIV.2609.17523. URL <https://doi.org/10.48550/arXiv.2609.17523>.
 - [32] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *CoRR*, abs/2504.08066, 2025. doi: 10.48550/ARXIV.2504.08066. URL <https://doi.org/10.48550/arXiv.2504.08066>.
 - [33] Junlin Yang, Che Jiang, Yu Fu, Tianwei Luo, Can Ren, Weizhi Wang, Kaikai Zhao, Hongyi Liu, Yuxin Zuo, Yuru Wang, Yuchen Fan, Kai Tian, Zhenzhao Yuan, Xiaojian Lin, Li Sheng, Rushi Qiang, Guoli Jia, Xingtai Lv, Ermo Hua, Dianqiao Lei, Youbang Sun, Ning Ding, Bowen Zhou, and Kaiyan Zhang. Frontis-ma1: Training an AI4AI model towards recursive self-improvement in machine learning engineering. *CoRR*, abs/2607.28568, 2026. doi: 10.48550/ARXIV.2607.28568. URL <https://doi.org/10.48550/arXiv.2607.28568>.
 - [34] Ling Yang, Zhenfei Yin, and Yingcheng Wu. Discovery foundation models: Toward open-ended discovery intelligence. *CoRR*, abs/2609.15973, 2026. doi: 10.48550/ARXIV.2609.15973. URL <https://doi.org/10.48550/arXiv.2609.15973>.
 - [35] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

- [36] Haotian Ye, Haowei Lin, Jingyi Tang, Yizhen Luo, Rahul Thapa, Caiyin Yang, Chang Su, Rui Yang, Ruihua Liu, Rundao Li, Zeyu Li, Pengwei Sun, Chong Gao, Dachao Ding, Guangrong He, Miaolei Zhang, Lina Sun, Wenyang Wang, Yuchen Zhong, Zhuohao Shen, Puheng Li, Pan Lu, Bianxiao Cui, Di He, Jianzhu Ma, Junfeng Li, Hexi Baoyin, Yejin Choi, Stefano Ermon, Xiaowen Chu, Tongyang Li, Yuzhi Xu, and James Zou. Structured scaling of ai discovery across diverse scientific domains. arXiv:2604.19341v2, 2026. URL <https://arxiv.org/html/2604.19341v2>.
- [37] Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-harness: Harnesses that improve themselves. CoRR, abs/2606.09498, 2026. doi: 10.48550/ARXIV.2606.09498. URL <https://doi.org/10.48550/arXiv.2606.09498>.
- [38] Jenny Zhang, Shengran Hu, Cong Lu, Robert T. Lange, and Jeff Clune. Darwin gödel machine: Open-ended evolution of self-improving agents. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aa5f5e6eb6f613ec412f1d948dfa21a5-Abstract-Conference.html.
- [39] Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob N. Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. CoRR, abs/2603.19461, 2026. doi: 10.48550/ARXIV.2603.19461. URL <https://doi.org/10.48550/arXiv.2603.19461>.
- [40] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y. Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a94ff6f922d995d7d3f4ebf4143e442-Abstract-Conference.html.
- [41] Yunwei Zhang, Qiaochu Tang, Yao Zhang, Jiabin Wang, Ulrich Stimming, and Alpha A. Lee. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nature Communications*, 11(1):1706, April 2020. doi: 10.1038/s41467-020-15235-7. URL <https://doi.org/10.1038/s41467-020-15235-7>.
- [42] Tong Zheng, Xidong Wu, Zheng Zhang, Zhankui He, Chaoyi Zhang, Benjamin Coleman, Ruojiao Wei, Di Bai, Haolin Liu, Rui Liu, Xue Wang, Yue Zhuan, Wang-Cheng Kang, Renkai Xiang, Heng Huang, Xinwu Cheng, and Yunsong Guo. Dream-rsi: Recursive self-improvement through evolving worlds. CoRR, abs/2609.14858, 2026. doi: 10.48550/ARXIV.2609.14858. URL <https://doi.org/10.48550/arXiv.2609.14858>.

DiscoveryRSI

Supplementary Methods and Evaluation

This supplement details the research workflow, hypothesis credit optimization, training implementation, and evaluation protocols supporting the main report. Appendices A–C describe the system and learning procedure; Appendices D–F describe automated assessment, human review, and literature-inspiration evaluation. Notation is defined where it is used. References to main-text figures and equations retain their original numbering.

Appendix Contents

Appendix A. Research Workflow and Linked Records	25
A.1. Research Objects and Agent Roles	25
A.2. Record Types and Provenance	25
A.3. From Contributions to Executable Investigations	26
A.4. Harness Replay and Rediscovery	26
Appendix B. Hypothesis Credit Optimization	26
B.1. Pairwise Preference Assessment	27
B.2. Hypothesis Weights	28
B.3. Contributor Shares and Provenance	29
B.4. Scientific-Family Credit Limits	29
B.5. Objective Balancing	30
B.6. Allocation Parameters	30
B.7. Credit Conservation	31
Appendix C. Training Data and Implementation	31
C.1. Trajectory Collection and Selection	32
C.2. Supervision and Optimization	32
C.3. Training Configuration	33
C.4. Serialization and Target Selection	33
C.5. Windowing: Preserved Targets and Conditional Equivalence	34
Appendix D. DFM-Bench Evaluation Protocol	34
D.1. Hypothesis Quality and Automated Review	34
D.2. Revisions, Development Paths, and Research Runs	35
D.3. Scientific Contribution, Delivery, and Calibration	36
D.4. Development-Path and Research-Run Rubrics	36
D.5. Assessment Records and Missing Evidence	37
D.6. Benchmark Coverage Definitions	37
Appendix E. Human Assessment Protocol	38
E.1. Review Cohort and Presentation	38
E.2. Rating Scales and Recommendations	38
E.3. Revision Outcomes and Interpretation	38
Appendix F. Literature-Inspiration Evaluation	39
F.1. Reference Corpus and Temporal Boundaries	39
F.2. Retrieval and Automated Judgment	39
F.3. Population Accounting	40
F.4. Claude Opus 5: Retrieval Sensitivity	40

Appendix A. Research Workflow and Linked Records

DiscoveryRSI develops hypotheses comprising a scientific claim, investigation plan, decision criteria, and alternative explanations. This appendix describes the roles and records connecting hypothesis development to task delivery and subsequent learning.

A.1. Research Objects and Agent Roles

A *hypothesis version* is one recorded proposal; a *revision* links it to its predecessor. Crossover can introduce several parents, so the *research tree* is generally a directed acyclic graph. A *development path* follows ordered hypothesis versions; a *research run* contains the complete graph and its decisions. An *agent trajectory* records one agent invocation, including instructions, actions, and observations. Development paths and agent trajectories are distinct evaluation and training units.

Lead chooses directions, parents, operations, and selections. Generation proposes hypotheses; Evolution develops or repairs them. Critique provides blinded preferences, Progress assesses revisions against ancestry, and Compiler translates selected plans into Harbor tasks. Evolve and repair retain an existing direction, crossover combines directions, and restart opens a new question. Shared context can inform a restart without creating parentage.

A.2. Record Types and Provenance

The record layer links hypotheses to evidence, data, contributions, and execution. Ancestry records scientific development; evidence reuse and execution dependencies are separate relations. Table 2 summarizes the retained records.

Table 2 | Persistent records supporting research and recursive improvement. Files retain material; explicit references retain provenance and relationships.

Record type	Contents	Recorded relationships
Hypothesis versions	Claim, plan, criteria, alternatives, and parents.	Revision ancestry; restart context recorded separately.
Literature and evidence	Sources, excerpts, and retrieval metadata.	Origin, dates, versions, and content hashes.
Data dependencies	Files and inputs required for investigation.	Bindings to hypotheses, jobs, and tasks.
Supporting artifacts	Scripts, results, plots, and evidence notes.	Explicit hypothesis attachments with paths and hashes.
Drafts and notes	Working proposals and continuation notes.	Agent ownership and recorded publication links.
Scientific assessments	Validity, preferences, and Progress judgments.	Assessed objects, comparison pairs, and ancestors.
Contributions and decisions	Assignments, roles, and publication events.	Agent, episode, attempt, and work identifiers.
Execution and delivery	Jobs, outputs, packages, and validation results.	Input dependencies and produced artifacts.

Registry entries and filesystem artifacts implement these relationships. Source records retain retrieval time and available publication metadata separately. A provenance link establishes recorded use or production; scientific assessment determines how the material supports or limits a claim.

A.3. From Contributions to Executable Investigations

Credit preparation links an assessed hypothesis to producing, supporting, and coordinating agent trajectories. Publication identity, original drafts, artifact hashes, and event order distinguish original work from later quotation or reuse. These links supply the contributor pools in Equation (18). Unresolved associations receive no allocated share.

For example, Generation may propose a comparison, Evolution may revise it after inspecting evidence, and Lead may supply the assignment that enabled that work. Their records can identify supported actions across several trajectories. Contribution allocation determines each trajectory’s weight; target selection determines which of its assistant outputs receive supervision (Appendices B and C).

A selected computational investigation is compiled into a Harbor task containing its inputs, environment, required outputs, and verification checks. Compilation preserves the population or biological system, comparisons, discriminating outcomes, and decision criteria. Unavailable measurements and wet-lab experiments remain explicit dependencies. Compilation and readiness checks assess the computational package; completed execution and scientific assessment establish what evidence the investigation supplies.

A.4. Harness Replay and Rediscovery

DiscoveryRSI uses Codex and Claude Code for sessions, context management, filesystem access, and tool execution [2, 22]. Research adapters connect these facilities to agent roles, shared records, computation, and task delivery. Harness changes concern coordination, evidence access, review timing, and research allocation.

Diagnosis and replay. Diagnostic runs hold the checkpoint and harness fixed. Linked decisions and execution records distinguish scientific-plan failures from tool failures, lost artifacts, and premature closure. Candidate repairs address evidence reuse, handoffs, retries, or completion checks; these examples do not constitute ablation results.

Replay preserves observations, tool access, and memory at the restored point. Alternatives require recorded continuations or new execution from a compatible state; ancestry alone supplies no counterfactual outcomes. This use of experience relates to Dream-RSI’s exploration-policy perspective [42].

Acceptance and renewed investigation. Harness comparisons hold the checkpoint, inputs, data access, and resource limits fixed. Model and harness candidates are assessed separately and jointly before deployment on fresh investigations (Algorithm 1).

Appendix B. Hypothesis Credit Optimization

Hypothesis credit optimization converts scientific assessments into weighted supervision for documented research contributions. The procedure first estimates relative hypothesis

preference, then assigns hypothesis weights, allocates contributor shares, limits credit within scientific families, and balances the training objective. Bayesian preference inference and contribution allocation are distinct steps.

B.1. Pairwise Preference Assessment

The preference model summarizes blinded pairwise Critique outcomes under a common scientific rubric. Its posterior describes relative preference strength; contributor attribution is determined separately from provenance.

The runtime collects blinded pairwise Critique outcomes: preference for A or B, a tie, or insufficient evidence. To derive the preference posterior, treat each versioned hypothesis as the rated entity in TrueSkill [10]. Its latent quantity z_h represents preference strength under the scientific critique rubric, rather than an actor’s personal skill. For one compatible comparison population, take Gaussian priors and noisy comparison performances:

$$z_h \sim \mathcal{N}(\mu_{h,0}, \sigma_{h,0}^2), \quad t_{h,e} = z_h + \epsilon_{h,e}, \quad \epsilon_{h,e} \sim \mathcal{N}(0, \beta^2). \quad (8)$$

For comparison $e = (a, b, o_e)$, the observed outcome depends on $D_e = t_{a,e} - t_{b,e}$: A is preferred when $D_e > \epsilon$, B when $D_e < -\epsilon$, and a tie when $|D_e| \leq \epsilon$. Here β models performance noise and ϵ is the draw margin. With $\Delta z = z_a - z_b$ and standard normal CDF Φ , integrating out performance noise gives

$$\begin{aligned} P(A \mid z_a, z_b) &= \Phi\left(\frac{\Delta z - \epsilon}{\sqrt{2}\beta}\right), \\ P(B \mid z_a, z_b) &= \Phi\left(\frac{-\Delta z - \epsilon}{\sqrt{2}\beta}\right), \\ P(\text{tie} \mid z_a, z_b) &= \Phi\left(\frac{\epsilon - \Delta z}{\sqrt{2}\beta}\right) - \Phi\left(\frac{-\epsilon - \Delta z}{\sqrt{2}\beta}\right). \end{aligned} \quad (9)$$

Insufficient evidence is an abstention and supplies no preference-likelihood factor. Under the comparison model’s conditional-independence assumption, Bayes’ rule yields the joint posterior

$$p(\mathbf{z} \mid \mathcal{D}_{\text{cmp}}) = \frac{1}{Z} \prod_h \mathcal{N}(z_h; \mu_{h,0}, \sigma_{h,0}^2) \prod_{e \in \mathcal{D}_{\text{cmp}}} P(o_e \mid z_{a_e}, z_{b_e}), \quad (10)$$

where Z is the evidence normalizer. Pairwise factors couple hypotheses, so this joint posterior is generally neither Gaussian nor factorized. TrueSkill uses approximate message passing and Gaussian moment matching to retain tractable marginal beliefs. Stored online updates summarize the evidence available at their update time; they are reused without claiming an exact batch re-inference of the complete comparison graph.

How a comparison changes the posterior. For a single decisive A-preferred comparison with independent Gaussian incoming beliefs, let

$$m = \mu_a - \mu_b, \quad c^2 = \sigma_a^2 + \sigma_b^2 + 2\beta^2, \quad u = \frac{m - \epsilon}{c}, \quad V(u) = \frac{\phi(u)}{\Phi(u)}, \quad W(u) = V(u)[V(u) + u], \quad (11)$$

with standard normal density ϕ . Before observing the outcome, $D \sim \mathcal{N}(m, c^2)$. Conditioning on $D > \varepsilon$ truncates this Gaussian, giving $\mathbb{E}[D | A] = m + cV(u)$ and $\text{Var}(D | A) = c^2[1 - W(u)]$. Gaussian conditioning transfers these moments to the latent hypotheses:

$$\mu'_a = \mu_a + \frac{\sigma_a^2}{c}V(u), \quad \mu'_b = \mu_b - \frac{\sigma_b^2}{c}V(u), \quad (12)$$

$$(\sigma'_a)^2 = \sigma_a^2 \left[1 - \frac{\sigma_a^2}{c^2}W(u) \right], \quad (\sigma'_b)^2 = \sigma_b^2 \left[1 - \frac{\sigma_b^2}{c^2}W(u) \right]. \quad (13)$$

These are the exact marginal first two moments for this single Gaussian-input comparison; retaining Gaussian beliefs is the approximation. A tie instead uses a two-sided truncated Gaussian. In a larger or sequential comparison graph, message schedules and Gaussian projection determine the stored approximation. Versioned runtime priors, noise, draw margin, and any dynamics variance remain part of the rating provenance; the static derivation does not silently replace those settings.

The stored marginals summarize relative preference strength within compatible source-snapshot, dataset, rubric, and connected comparison strata. Reviewer bias, scientific uncertainty, and dependence between repeated critiques are not fully represented by the Gaussian variance. Separate ancestor-aware Progress judgments supply transition evidence.

B.2. Hypothesis Weights

For a rated hypothesis, define a conservative ranking statistic and its tie-aware within-stratum rank:

$$s_h = \mu_h - \kappa\sigma_h, \quad (14)$$

$$r_h = \frac{\sum_{k \in \mathcal{S}} \mathbf{1}[s_k < s_h] + \frac{1}{2} \sum_{k \in \mathcal{S}} \mathbf{1}[s_k = s_h]}{|\mathcal{S}|}. \quad (15)$$

Under the Gaussian marginal approximation, s_h is the $\Phi(-\kappa)$ posterior quantile; for $\kappa = 2$, approximately 2.3% of the modeled preference mass lies below it. This is a conservative statistic of rubric-conditioned preference. A singleton has rank 1/2, supplying no relative ordering. Let n_h be the effective comparison count. Two reliability factors damp the contribution of sparse or uncertain ratings:

$$\rho_h^{\text{cov}} = \frac{n_h}{n_h + \nu}, \quad \rho_h^{\text{unc}} = \frac{1}{1 + \sigma_h/\sigma_{\text{ref}}}. \quad (16)$$

Let ν_h indicate a recorded valid assessment and at least one effective comparison, with no recorded marginal-repetition judgment. Let p_h indicate a recorded substantive-progress or necessary-repair judgment. The bounded outcome transformation is given in Equation (17). Validity here means a coherent, assessable research design under the stored rubric, not experimental confirmation. The gate controls eligibility for this weak-supervision route. A missing rating or assessment remains an unknown outcome; zero assigned loss is not a negative scientific label. A valid hypothesis motivated by a supported null result remains eligible.

$$q_h = \nu_h \min\{q_{\max}, q_{\min} + (q_{\max} - q_{\min})r_h\rho_h^{\text{cov}}\rho_h^{\text{unc}} + \lambda_p p_h\}. \quad (17)$$

Equation (17) is our bounded transformation of a preference posterior, not a quantity inferred by TrueSkill. The ranking statistic and uncertainty factor both penalize uncertainty; comparison count also correlates with posterior concentration. Their individual effects require ablation. A Progress bonus supplements the outcome weight without changing the TrueSkill posterior.

B.3. Contributor Shares and Provenance

We construct an auditable contribution graph whose edges identify an agent invocation’s relationship to a hypothesis. Direct publication requires matching publisher, episode, and creation-time provenance. Documented production or support can additionally be recovered from original prepublication proposal text, an exact draft-revision promotion record, or emitted supporting-file bytes whose hash and original path match the frozen candidate record. Lead coordination requires a contributing committed work assignment or publication/promotion record. Merely sharing an episode, appearing in an ancestry graph, or completing a worker invocation creates no credit edge. These links establish recorded association and production provenance; they do not establish sole authorship or the artifact’s scientific correctness.

For hypothesis h , let L_h contain verified coordination trajectories and P_h contain direct production or documented supporting trajectories. Each trajectory belongs to at most one pool for a given hypothesis. With coordination fraction γ , define

$$\alpha_{ih} = \begin{cases} \gamma/|L_h|, & i \in L_h, \\ (1 - \gamma)/|P_h|, & i \in P_h, \\ 0, & \text{otherwise.} \end{cases} \quad \sum_i \alpha_{ih} \leq 1. \quad (18)$$

Only nonempty pools are evaluated. If a contributor pool is missing, its share remains unallocated rather than being transferred to a convenient recipient. A Lead-only publication with unresolved authorship therefore receives the coordination share alone. Conversely, documented Lead production can enter P_h : the rule follows contribution function rather than a fixed role label.

For example, one coordinating Lead, one Generation publisher, and one Evolution trajectory with documented supporting work receive fractions 0.2, 0.4, and 0.4 under the implemented $\gamma = 0.2$ policy. This allocation reaches work that need not register its own hypothesis. However, the fractions are a fixed responsibility prior, not a Bayesian posterior over actors or a counterfactual estimate of their marginal contribution. The TrueSkill marginal informs q_h through our explicit transformation; provenance supplies α_{ih} . Critique votes cannot independently certify the correctness of those same critique outputs, and successful compilation cannot certify scientific fidelity. Such roles require their own evidence before receiving direct imitation targets.

B.4. Scientific-Family Credit Limits

Duplicate registrations do not receive additional scientific credit. We suppress exact duplicate proposal-credit events within a source snapshot and dataset. Local ancestry

families $F_{\text{loc}}(h)$ also receive a bounded outcome budget:

$$\tilde{q}_h = q_h \min \left\{ 1, \frac{C_{\text{loc}}}{\sum_{k \in F_{\text{loc}}(h)} q_k} \right\}, \quad u_{ih} = \alpha_{ih} \tilde{q}_h. \quad (19)$$

The local sum covers rated comparison nodes, including imported anchors when present; only generated hypotheses with contributor evidence allocate training credit. Global scientific families F_{glob} join recorded ancestry and compatible candidate or exact-proposal identities across snapshots. Their allocated credit is capped before forming a trajectory weight:

$$d_F = \min \left\{ 1, \frac{C_{\text{glob}}}{\sum_{h \in F} \sum_i u_{ih}} \right\}. \quad (20)$$

The final allocation from hypothesis h to trajectory i is $b_{ih} = d_{F_{\text{glob}}(h)} u_{ih}$, and the trajectory weight is $w_i = \min(W_{\text{max}}, \sum_h b_{ih})$. For a local family with zero total weight, the local cap factor is defined as one. For a zero-mass family, the cap factor is defined as one and allocated credit remains zero. These budgets limit the credit assigned to repeated versions of the same scientific question. They are pre-normalization credit constraints; subsequent domain and group balancing determines their relative influence in the final objective. Unresolved semantic duplication remains an audit limitation rather than an assumed solved problem.

B.5. Objective Balancing

Let g_i identify a conservative shared-memory group, closed over campaign memory, ancestry, resumed trajectories, and recovered duplicates; these balancing groups are distinct from the scientific credit families above. For domain d , let $\mathcal{G}_d$ contain groups with retained targets and let η_d be its desired objective mass. The fixed trajectory coefficient and intended update are

$$\omega_i = \frac{\eta_{d_i}}{|\mathcal{G}_{d_i}|} \frac{w_i}{\sum_{k: d_k=d_i, g_k=g_i} w_k}, \quad \sum_d \eta_d = 1. \quad (21)$$

All sums in Equation (21) concern retained trajectories with positive target mass. Coefficients and denominators are fixed from the frozen training release and are not renormalized within individual minibatches. This prevents transcript length alone from determining a trajectory’s loss weight. The coefficients allocate objective mass, not guaranteed shares of gradient norm. Initialization and optimization settings are reported in Appendix C.

B.6. Allocation Parameters

Quantity	Value	Function
Conservative-rating factor κ	2	Penalize uncertain preference means
Coverage constant ν	2	Damp sparsely compared hypotheses
Uncertainty reference σ_{ref}	8.3333	Fixed rating-scale reference
Weight interval $[q_{\text{min}}, q_{\text{max}}]$	[0.5, 2]	Bound eligible outcome weights
Progress supplement λ_p	0.25	Use recorded progress evidence

Quantity	Value	Function
Coordination fraction γ	0.2	Reserve coordination share
Local/global family budgets	2, 2	Limit repeated scientific credit
Trajectory cap $W_{\max}$	2	Bound aggregated trajectory weight
Positive local credit c_{it}	0.5	Uniform positive token eligibility

These constants specify the implemented allocation recipe; they are not estimates of causal contribution or empirically optimal hyperparameters. The uniform local credit cancels in the per-trajectory mean. Domain balancing is separate from scientific-quality ranking; a percentile within one stratum is not a universal cross-domain utility scale.

B.7. Credit Conservation

Proposition 1 (Pre-normalization credit conservation) *Fix deduplicated hypothesis-credit events and scientific-family assignments. Assume nonnegative hypothesis weights, positive budgets, and contributor shares satisfying $\sum_i \alpha_{ih} \leq 1$. Under the local and global scaling rules,*

$$\sum_i b_{ih} \leq \tilde{q}_h, \quad \sum_{h \in F} \sum_i b_{ih} \leq C_{\text{glob}}, \quad 0 \leq w_i \leq W_{\max}$$

for each global scientific family F . Missing contributor shares remain unallocated.

Proof. Fix deduplicated credit events and a partition into global scientific families, with nonnegative weights, positive budgets, and $0 \leq \gamma \leq 1$. Every cap factor belongs to $[0, 1]$. The mutually exclusive contributor pools give $\sum_i \alpha_{ih} \leq 1$, hence

$$\sum_i b_{ih} = d_{F_{\text{glob}}(h)} \tilde{q}_h \sum_i \alpha_{ih} \leq \tilde{q}_h. \quad (22)$$

For a global family F , let $U_F = \sum_{h \in F, i} u_{ih}$. If $U_F > 0$, then

$$\sum_{h \in F, i} b_{ih} = \min(1, C_{\text{glob}}/U_F) U_F = \min(U_F, C_{\text{glob}}) \leq C_{\text{glob}}. \quad (23)$$

If $U_F = 0$, all its allocations vanish. Similarly, the local scaling gives $\sum_{h \in F_{\text{loc}}} \tilde{q}_h \leq C_{\text{loc}}$ for each fixed local family. Nonnegativity and the trajectory cap yield $0 \leq w_i \leq W_{\max}$. Missing contributor pools do not enter α_{ih} and their reserved shares remain unallocated. This proves [Proposition 1](#). The result holds before objective balancing; group normalization can change relative family influence and supplies no corresponding bound on gradient magnitude.

Appendix C. Training Data and Implementation

This appendix documents the data and optimization used for full-parameter fine-tuning of DiscoveryRSI-27B. An agent trajectory is one recorded agent invocation, rather than a path through the hypothesis tree. Only supported assistant outputs receive direct loss; other retained tokens provide context.

C.1. Trajectory Collection and Selection

The source collection contains 15,792 research trajectories. Structural review identifies 13,270 reviewable agent trajectories, of which 13,269 pass export-integrity verification. Attribution identifies 2,432 agent trajectories with eligible supervision targets. Native-template tokenization retains 2,397 agent trajectories; 35 agent trajectories with unsupported mid-conversation developer messages remain excluded from the training export.

Table 4 | Trajectory and token accounting for hypothesis-credit training. Training-sequence tokens include supervised targets and zero-loss context. Collected trajectories and retained training trajectories are counted separately.

Role	Collected	Training	Sequence tokens	Supervised tokens
Lead	3,423	634	98,357,399	977,931
Generation	2,502	861	96,414,270	2,437,985
Evolution	2,310	902	96,349,331	2,761,689
Critique	4,988	0	—	0
Compiler	2,056	0	—	0
Progress	513	0	—	0
Total	15,792	2,397	291,121,000	6,177,605

Lead, Generation, and Evolution account for 26.4%, 35.9%, and 37.6% of retained training trajectories, respectively. Critique, Compiler, and Progress records remain available as sources or context but receive no direct supervision. Exclusion from supervision indicates insufficient support under the attribution policy and does not imply that a role or trajectory lacks scientific value.

C.2. Supervision and Optimization

Scientific assessments and documented contributions determine fixed trajectory weights. Supervision targets supported assistant responses and tool-call arguments. Tool observations, failed actions, repeated context, and unsupported actions receive zero direct loss. For trajectory i , let ω_i denote its fixed weight, c_{it} the supervision credit assigned to token t , and $M_i = \sum_t c_{it}$ its supervised mass. The implemented objective is

$$\mathcal{L}_{\text{credit}}(\theta) = - \sum_i \frac{\omega_i}{M_i} \sum_t c_{it} \log \pi_{\theta}(a_{it} \mid x_{it}). \quad (24)$$

Weights and normalization factors are frozen before optimization. The objective uses offline weighted supervision; it does not involve online policy optimization.

The maximum training window is 262,144 tokens. Seven trajectories require multiple windows, producing 2,404 windows from 2,397 unique trajectories. Windows preserve canonical turn and tool-batch boundaries. Repeated initial and left context receives zero loss, and all windows from a trajectory share its supervision denominator and optimizer batch. The retained trajectories contain 304,424,899 tokens before window preparation. The final training sequences contain 291,121,000 tokens and preserve all 6,177,605

supervised targets. These totals refer to different preparation stages; the main text reports the final training-sequence count.

C.3. Training Configuration

Training uses full-parameter language-model fine-tuning with BF16 precision, Adam, a learning rate of 5×10^{-6} , and a global batch size of 16. One epoch completes in 150 optimizer steps. The distributed configuration uses 64 GPUs with tensor parallelism 4, context parallelism 8, and data parallelism 2. The final batch is padded with zero-loss entries to preserve the nominal batch normalization.

C.4. Serialization and Target Selection

Each trajectory retains its source identifier, original messages and tool definitions, linked hypotheses, contribution evidence, and exposure metadata. A token mask records eligible original assistant outputs and zero-loss positions. Instructions and tool observations remain context; known failed actions, repeated context, and unavailable reasoning placeholders receive zero direct loss. Serialization preserves instruction order and excludes future information from earlier contexts. Unsupported instruction sequences are omitted from the training export while their source records remain available.

The data preparation groups shared memory, ancestry, resumed interactions, and recovered duplicates before assigning balancing coefficients. These groups are distinct from the scientific families used to cap hypothesis credit. Targets are counted once across windows, and each trajectory retains one denominator. The fixed token coefficient under uniform trajectory sampling is $N\omega_i/M_i$, where $N = 2,397$ and M_i is the trajectory’s total supervised mass. Zero-loss padding preserves the nominal batch size at the end of the epoch.

The implemented token credit is constant at $c_{it} = 0.5$ on eligible targets and zero elsewhere. This positive constant cancels between the numerator and M_i , giving the eligible-token mean in Eq. (4). Thus the main-text objective and Eq. (24) describe the same normalization. The weights specify relative loss contributions, not guaranteed shares of gradient magnitude.

Preparation checks cover source integrity, token/mask alignment, unique target coverage, fixed trajectory denominators, and domain-weight accounting. They validate the training representation; scientific correctness is assessed separately. In particular, conserving supervised targets and credit does not establish that the underlying hypothesis is scientifically supported. The source records and private research artifacts are not distributed with this supplement.

C.5. Windowing: Preserved Targets and Conditional Equivalence

Proposition 2 (Objective preservation under windowing) *Let U_i be the eligible target positions of trajectory i , with $|U_i| > 0$, and let $\{U_{iw}\}_w$ partition U_i . If each target retains its original conditioning context, all windows use the same fixed coefficient ω_i , and normalization uses $|U_i|$, then*

$$-\sum_i \frac{\omega_i}{|U_i|} \sum_w \sum_{u \in U_{iw}} \log \pi_\theta(a_{iu} \mid x_{iu}) = -\sum_i \frac{\omega_i}{|U_i|} \sum_{u \in U_i} \log \pi_\theta(a_{iu} \mid x_{iu}).$$

The windowed and original objectives therefore have identical gradients wherever defined.

Proof and implementation scope. The target sets U_{iw} partition U_i , so each target appears exactly once. With identical conditioning contexts, summing their log-likelihoods over windows gives the full-trajectory sum. Multiplication by the fixed coefficient $-\omega_i/|U_i|$ proves equality of the objectives and their gradients. Repeated context has zero loss. A constant positive token credit cancels in the numerator and denominator.

The conditioning assumption is stronger than target conservation. The implemented 256K export retains initial instructions and bounded overlap, so a window can omit earlier context from a longer trajectory. It preserves target counts and fixed normalization but does not generally preserve full-history likelihoods or gradients. The proposition establishes equivalence only when conditioning is unchanged; independently renormalizing each window would also violate its assumptions.

Appendix D. DFM-Bench Evaluation Protocol

DFM-Bench evaluates five objects: individual hypotheses, revisions, development paths, research runs, and scientific contributions. This appendix defines their scoring, evidence requirements, and coverage. The version of the rubric and evaluator is retained with each assessment; results from different applicability rules are not pooled without identifying those differences.

D.1. Hypothesis Quality and Automated Review

The hypothesis profile uses eight dimensions: importance, novelty, grounding, falsifiability, data resolvability, feasibility, alternatives, and actionability. Their weights are 20, 20, 15, 10, 10, 10, 10, and 5, respectively, on a 1–10 scale. These are frozen internal protocol conventions, not empirically established optimal weights.

A rubric judge, a domain-scientist role, and a methods-reviewer role assess the same evidence independently of one another’s grades. The current panel uses separate prompts to the same model family. It therefore measures correlated AI perspectives, not three independent expert opinions. Each dimension takes the median when at least two reviewers provide a known score. A complete quality score Q requires all three reviews, majority eligibility, and all eight dimension medians. The dimension weights sum to 100 and define a weighted mean on the 1–10 scale. Unknown evidence remains unknown rather than becoming zero.

Review packets use source-verbatim cards and traceable supporting material. They distinguish a proposal’s scientific importance from its maturity: an important question need

Table 5 | The five DFM-Bench evaluation objects. Scientific quality is evaluated at each level using evidence appropriate to that object. Delivery maturity is a separate axis.

Level	Object	Core scientific question	Outputs
Hypothesis	Frozen candidate plus evidence	Is this a worthwhile, grounded, testable research opportunity?	Eight-dimension profile; Q /10; unknowns and review coverage.
Revision	Actual parent-child pair	Did this step advance understanding, repair an important defect, or repeat prior work?	Substantive progress, necessary repair, repetition, insufficient evidence; numeric ΔQ separately.
Development path	Ordered hypothesis path	Does inquiry accumulate evidence and revise decisions coherently?	Seven-axis L1-L7 profile; complete, scope-matched composite /100.
Research run	Full research graph and decisions	Does the portfolio address meaningful opportunities and allocate effort well?	Independent T1-T7 profile; coverage, synthesis, allocation, and frontier diagnostics.
Scientific contribution	Deduplicated contribution family	What question, distinction, evidence, negative result, or uncertainty resolution was added?	Contribution ledger, evidence maturity, novelty scope, and adjudication status.

not already have a positive result. Conversely, a completed script does not demonstrate a consequential finding. The packet’s available literature bounds novelty judgments; global novelty is not inferred from limited retrieval.

D.2. Revisions, Development Paths, and Research Runs

Revision evaluation retains the four categorical judgments used by the progress contract. Substantive-progress rates, useful-step rates including repairs, repetition, and insufficient-evidence coverage are reported separately. Numeric $\Delta Q = Q_{\text{child}} - Q_{\text{parent}}$ is defined when both automated quality scores are available.

Development-path evaluation assesses explicitly ordered paths. The recorded protocol selects one longest-depth path to each terminal node, breaking ties by identifier, and excludes singleton paths. At crossovers, off-path parents remain context. Coverage therefore counts evaluated paths rather than every possible route through the graph. The seven weights are 5/30/20/20/10/10/5. Research-run evaluation independently inspects the entire graph using weights 25/15/15/20/10/10/5. Both use 0–4 dimension anchors, converted to /100 only under the required completeness and applicability rules.

A legitimately inapplicable dimension differs from an unknown one. Restricted-scope composites are grouped by applicability mask; missing required evidence prevents a full composite. Research-run scores are not averages of development-path scores. Shared ancestors also make path-level measurements correlated, precluding naive independent-sample interpretations.

D.3. Scientific Contribution, Delivery, and Calibration

Scientific-contribution assessment groups related candidates into contribution families and records what was added beyond the starting evidence. Family membership does not imply confirmation; maturity cannot transfer automatically from one member to all others. Negative or inconclusive evidence can matter when it changes an important scientific decision. Human adjudication and independent evidence are needed for stronger claims of novelty and discovery.

Delivery is recorded alongside this scientific profile: proposed, selected, compiled, readiness-passed, executed, and independently reviewed. An unattempted check is distinct from a failure. The protocol does not combine delivery maturity and scientific scores into one overall number.

Calibration concerns the meaning and consistency of judgments. The implemented workflow freezes domain anchors, packets, aggregation, disagreements, and missingness. Same-packet comparisons against fresh-context AI references provide development diagnostics. Human calibration requires independent expert judgments, adjudication, and unseen acceptance cases; it remains pending. Prompt-role agreement alone does not establish validity.

The DFM framework’s distinction between current knowledge progress and future capability growth [34] applies across all five objects. A good final hypothesis does not establish a sound development path, and a disciplined research run does not establish that the trained model improves on future research runs. Frontier plots are useful diagnostics, but a retrospective maximum over final-snapshot scores is monotone by construction and cannot demonstrate learning.

D.4. Development-Path and Research-Run Rubrics

Dimension scores use 0–4 evidence anchors: absent or contradicted, recognized with major gaps, partly substantiated, consistently substantiated, and comprehensive within the declared scope. Domain-specific evidence requirements qualify these anchors. The following labels summarize the clinical rubric; weights are fixed protocol choices rather than calibrated universal utilities.

Dimension	Weight	Criterion
L1	5%	Relevant disciplinary connections
L2	30%	Technical and knowledge exploration
L3	20%	Cumulative scientific progress
L4	20%	Evidential and inferential discipline
L5	10%	Scientific continuity and synthesis
L6	10%	Adaptive continuation and stopping
L7	5%	Efficient use of the path’s budget

Dimension	Weight	Criterion
T1	25%	Disciplinary coverage and integration
T2	15%	Technical and knowledge exploration
T3	15%	Cumulative scientific progress
T4	20%	Evidential and inferential discipline
T5	10%	Portfolio synthesis
T6	10%	Adaptive scientific decisions
T7	5%	Budget-aware exploration

D.5. Assessment Records and Missing Evidence

Each assessment identifies its object, source snapshot, rubric and protocol versions, supporting evidence, scores, rationale, and review status. Hypothesis assessments use a frozen proposal and its evidence. Revisions require actual parent-child pairs; development paths require ordered ancestry; research-run assessment requires the full graph and recorded decisions. Agent trajectories supply evidence about the work behind these objects.

Scores are recorded as known, unknown, or inapplicable under the rubric. Unknown values are not replaced by zero. An incomplete assessment is reported with its coverage and does not receive a complete composite. Resource-allocation judgments require recorded budgets and opportunity information; proposal counts do not reconstruct missing budgets. Contribution records additionally identify the scientific question, the increment beyond starting evidence, and its status as supported, contradicted, inconclusive, or unresolved. Evidence maturity distinguishes proposed, source-grounded, diagnostically examined, study-completed, and independently checked contributions.

D.6. Benchmark Coverage Definitions

[Table 1](#) compares explicit evaluation constructs, not the capabilities of agents that happen to run on a benchmark. Hypothesis merit concerns scientific value and testability; contribution maturity distinguishes proposed, examined, and independently supported contributions. Evidence use concerns grounding analyses or claims in available evidence. Revision assessment compares a change against a predecessor; development-path assessment examines cumulative inquiry; whole-run allocation assesses decisions among research directions. A cross means no explicit assessment under these definitions, not that the benchmark lacks scientific value. Partial coverage identifies a related measure with narrower scope.

AgentIdeaBench scores final ideas with literature-verified critics; its grounding assessment is partial evidence-use coverage. Its generation-time refinement experiments do not constitute an explicit revision or ancestry-path evaluation object [21]. DiscoveryBench evaluates reference-aligned hypotheses and analytical facets for supplied discovery goals; this provides partial hypothesis-merit and evidence-use coverage, rather than an open-ended merit or contribution-maturity rubric [20]. BLADE compares conceptual variables, transformations, and statistical models against expert analyses, providing partial evidence-use coverage through analytical decisions [8]. ScienceAgentBench

assesses generated programs, outputs, and task-stage rubrics; this is partial evidence-use coverage rather than evaluation of evolving scientific claims [3].

OEB explicitly assesses evidence, experiments, and responses to contradicted ideas. Its inquiry grouping and search-topology measures supply partial development-path coverage; its budget and concurrency profiles supply partial whole-run allocation coverage because they describe research style without ranking one style above another. It does not supply the proposal-merit or contribution-maturity assessments defined here [27]. DFM-Bench’s checks indicate implemented rubric coverage conditional on sufficient records, not independent expert validation or empirical superiority over these instruments.

Appendix E. Human Assessment Protocol

Human assessment examines the value and design quality of proposed investigations, with automated scores hidden. Overall value, criterion ratings, recommendations, and design-flaw verdicts are separate judgments.

E.1. Review Cohort and Presentation

The review involved 27 medical doctors, medical students, and bioscience PhD students with biomedical or clinical research backgrounds, under three senior supervisors. The fixed hypothesis-atlas cohort contains 292 proposals: HCC (52), MIMIC-IV (60), eICU (60), UK Biobank (55), GEO therapeutic targets (35), and GEO disease mechanisms (30). The groups span the four evaluated life-science domains.

English proposal packets included the scientific question, investigation plan, data requirements, alternative explanations, and citations. Model identity, execution provenance, and automated grades were hidden. Proposals were presented in randomized order, and reviewers assessed each packet without consulting its revision history. Figure 4 illustrates the interface.

E.2. Rating Scales and Recommendations

Reviewers first recorded overall research value on a five-point scale from not worth pursuing to top priority, together with a recommendation to prioritize, revise, defer, or reject. They then rated six criteria on anchored four-point scales: importance, advance over known work, plausible and evidence-bounded claims, data fit, alternative explanations, and informative results. A separate verdict classified consequential design flaws as absent, fixable, or fatal. Reviewers could select “Can’t judge” when evidence or expertise was insufficient; flaw checklists and explanatory comments were optional.

The five-point overall judgment is distinct from the four-point criterion profile. Recommendations to revise or defer do not imply endorsement, and design-flaw verdicts concern the investigation plan.

E.3. Revision Outcomes and Interpretation

The revision analysis compares human-rated quality for a revised hypothesis and its actual predecessor. Positive, zero, and negative score changes define improved, un-

changed, and worsened revisions, respectively. The reported mean change and outcome proportions characterize the assessed parent–child pairs. They do not estimate the effect of a model update across independent investigations.

Figure 8 separates criterion profiles and revision outcomes. The MIMIC-IV case (Figure 9) shows six-criterion means along a selected path; it does not imply that all branches improve. Shared ancestry makes revision observations dependent. This assessment is distinct from calibration of the automated evaluator against independently adjudicated, unseen cases.

Appendix F. Literature-Inspiration Evaluation

This appendix defines the retrospective literature comparison used in RQ4. “Inspiration” denotes a concrete scientific connection: a reference after the recorded cutoff provides an experimental design, comparator, measurement, or testable evidence gap relevant to developing or testing a generated hypothesis. It does not imply that the publication’s authors encountered the hypothesis or were causally influenced by it.

F.1. Reference Corpus and Temporal Boundaries

One model-blind reference corpus was constructed from deterministic source-only anchors. A completed citation wave followed 420 backward and forward citation traversals, yielding 17,790 citation edges and 16,440 resolved works. The same corpus was available to all model groups within the publication window from 1 April 2025 through 14 October 2026.

References were filtered by each group’s recorded operational cutoff. Exact-day boundaries were retained for GPT serving identifiers. Month-level boundaries were used for Qwen (April 2025, a provisional family proxy), Claude Opus 5 (May 2026), DeepSeek V4.1 (September 2026, a release-month proxy), and GLM-5.3 (August 2026, a release-month proxy). These boundaries make the comparison reproducible; they do not identify internal model knowledge cutoffs.

F.2. Retrieval and Automated Judgment

Candidate retrieval required the best reference card to share at least three non-stopword terms with the hypothesis. This lexical gate determines retrieval eligibility, not semantic relevance. GPT-6 Luna supplied the semantic judgments, with 145 successful judgments from 149 sampled records. Four records lacked successful judgments. Exact duplicate hypothesis–reference inputs shared a judgment, which was propagated to 903 registered hypothesis versions. The resulting labels are correlated automated assessments rather than independent expert reviews.

The inspiration rate uses all generated hypothesis versions as its denominator, as defined in Eq. (7). Only versions receiving a positive judgment contribute to the numerator. Consequently, absence of a positive label may reflect retrieval or audit coverage as well as lack of a qualifying connection. The underlying audit calls this endpoint reference actionability; the report uses the operational term inspiration rate consistently.

F.3. Population Accounting

DiscoveryRSI-27B pools two production runs. The overall numerator and denominator are 74/1,176; domain counts are shown below. The comparison groups are DeepSeek (14/257), GLM (1/36), GPT (59/2,341), and Claude (3/131). Qwen original remains in the underlying audit but is omitted from the main presentation figure.

Table 6 | DiscoveryRSI-27B literature-inspiration counts. Numerators count hypothesis versions with positive judgments; denominators include all generated versions.

Domain	Positive	Generated	Rate
Clinical and population health	13	557	2.3%
Therapeutic target prioritization	30	178	16.9%
Disease mechanisms	17	251	6.8%
Population multi-omics	14	190	7.4%
Total	74	1,176	6.3%

Group sizes and domain composition differ, and some strata are small. Versions from the same research family and reused judgments are dependent. The report therefore presents descriptive percentages without significance tests.

F.4. Claude Opus 5: Retrieval Sensitivity

The targeted analysis in RQ4 separates retrieval coverage from positive judgments conditional on retrieval. Only 15/131 Claude proposals passed the retrieval gate (11.5%), compared with 355/1,176 DiscoveryRSI proposals (30.2%). Among retrieved proposals, the positive fractions were similar: 20.0% and 20.8%, respectively.

Claude’s median proposal length was 41,134 characters, whereas retrieval used the first 6,000 characters and semantic review used the first 2,500. Provenance and revision notes could therefore displace scientific content. A diagnostic substituted proposal titles as queries while preserving the corpus, scoring method, date rule, and overlap threshold. Claude’s retrieval count increased from 15 to 95. This establishes sensitivity to representation, not additional positive semantic matches. A structured scientific representation and renewed judgments across all groups would be needed to assess the effect on the final comparison.

The audit observes literature correspondence, not researcher exposure, adoption, study initiation, or causal influence. Prospective evaluation would freeze generated hypotheses before researcher review and follow subsequent investigation plans, experiments, and publications.