

---

# ResearchGym: Evaluating AI Agents in the Full Scientific Inquiry Loop

---

Anonymous Authors<sup>1</sup>

## Abstract

While LLM-based agents are increasingly applied to scientific workflows , a fundamental question remains: Are current models truly qualified for the dynamic and uncertain process of scientific discovery? Existing evaluations, often limited to static tasks, fail to distinguish between genuine reasoning and rote memorization. In this work, we introduce *ResearchGym*, a diagnostic environment designed to probe the epistemic limits of agents by formalizing scientific inquiry into interactive "Research Trees". By representing research as directed acyclic graphs (DAGs) of logical dependencies, *ResearchGym* simulates the full scientific inquiry loop—from initial hypothesis formulation and study design to result interpretation and belief updating. Our evaluation across the RG-30 benchmark reveals two critical deficiencies. First, we identify the "Erosion of Marginal Capabilities": as agents engage in long-horizon interactions, they exhibit "cognitive tunneling," where their critical judgment and ability to detect anomalies degrade significantly compared to their intrinsic baselines. Second, we find a distinct boundary between interpolation and extrapolation: agent performance drops on papers published after their training data cutoff. This implies that their apparent competence is driven largely by parametric memory rather than the extrapolative reasoning required to explore the unknown. These findings highlight that scaling context alone is insufficient; achieving qualified AI scientists might require architectural innovations or strong human supervision to preserve critical oversight and generalization.

## 1. Introduction

The scientific enterprise is undergoing a phase transition as large language models (LLMs) and agentic scaffolds become tightly coupled with scientific workflows (Gottweis et al., 2025; Tie et al., 2025; Lu et al., 2024). This raises a central evaluation question: can current agents reliably participate in an *inquiry loop*, where they iteratively propose

a next step, receive feedback, and update beliefs and plans, rather than merely producing plausible one-shot outputs? Existing benchmarks provide valuable signals, but they typically cover only fragments of the research lifecycle, making it difficult to assess whether an agent can sustain coherent reasoning across the full inquiry loop; Table 1 illustrates this fragmentation, where ideation/planning benchmarks (e.g., IdeaBench and Idea2Plan) stop before execution and feedback (Guo et al., 2024; Huang et al., 2025b), while execution-centric benchmarks (e.g., ScienceAgentBench and SciCode) enable reliable scoring via code execution yet often assume the research goal is pre-specified and therefore do not test problem formulation or high-level scientific decision making (Chen et al., 2024; Tian et al.). Meanwhile, long-horizon agentic settings in ML (e.g., MLGym and MLE-Bench) emphasize iterative experimentation but can collapse discovery into metric optimization within a single domain (Nathani et al., 2025; Chan et al., 2024), and interactive simulators such as DiscoveryWorld move closer to end-to-end discovery but are often limited to relatively basic discovery settings with abstracted tools, leaving a gap to real-world frontier research discovery (Jansen et al., 2024).

At the same time, these benchmarks often operationalize the agent as an end-to-end "autonomous researcher" that produces a final artifact, whereas real scientific work is frequently conducted in an *inquiry style*: researchers iteratively propose the next step, observe feedback (including failures and constraints), request clarifications or hints, and update beliefs and plans. Motivated by both the lifecycle coverage gap and the need to evaluate this inquiry-loop interaction pattern, we propose *ResearchGym*, an inquiry-loop evaluation environment that models research as an interactive process rather than a single-shot task. *ResearchGym* formalizes a project as a *Research Tree* (a DAG of logical dependencies) and exposes the agent to a sequence of states (Topic/Subtopic/Study/Result), enabling multi-turn exploration, belief updating, and controlled feedback (including "Fake Results") within a unified framework. Further, starting from the neuroscience domain, we curate 30 papers from CNS-level venues to build **RG-30**, a benchmark suite of interactive environments that supports end-to-end evaluation across the inquiry loop.

**Contributions.** We make the following contributions:

	Benchmark	Subtopic Proposal	Study Design	Study Execution	Result Analysis
055	IdeaBench(Guo et al., 2024)	✓	✗	✗	✗
056	Idea2Plan(Huang et al., 2025b)	✗	✓	✗	✗
057	ScienceAgentBench(Chen et al., 2024)	✗	✗	✓	✗
058	SciCode(Tian et al.)	✗	✗	✓	✗
059	BAISBench(Luo et al., 2025a)	✗	✗	✓	✓
060	MLAgentBench(Huang et al., 2023)	✗	✗	✓	✓
061	MLGym-Bench(Nathani et al., 2025)	✗	✗	✓	✓
062	MLE-Bench(Chan et al., 2024)	✗	✗	✓	✓
063	BixBench(Mitchener et al., 2025)	✗	✗	✓	✓
064	RE-Bench(Wijk et al., 2024)	✗	✗	✓	✓
065	EXP-Bench(Kon et al., 2025)	✗	✗	✓	✓
066	PaperBench(Starace et al., 2025)	✗	✗	✓	✓
067	DiscoveryWorld(Jansen et al., 2024)	✓	✓	✓	✓
068	<b>RG-30</b> (Ours)	✓	✓	✗	✓

Table 1. Benchmark coverage across the scientific inquiry loop. We mark whether each benchmark involves key stages (subtopic proposal, study design, study execution, and result analysis). The table highlights that most prior work covers only a subset of the lifecycle.

- We introduce *ResearchGym*, an inquiry-loop evaluation environment that models scientific research as interactive state transitions rather than one-shot tasks.
- We formalize research episodes as *Research Trees* (DAGs of logical dependencies) extracted from scientific papers, enabling structured multi-turn exploration and belief updating.
- We build **RG-30**, a curated benchmark suite with hybrid evaluation to score both exploration behavior and evidence-grounded conclusions.
- We provide diagnostic analyses that reveal key failure modes of current models, including critical-judgment erosion under long-horizon interaction and a clear interpolation-extrapolation boundary.

## 2. Related Work

### 2.1. AI Agent for Science

Concurrently, agentic systems are being deployed across scientific domains. For example, Gottweis et al. introduce an AI co-scientist that iteratively generates and refines biomedical hypotheses in a multi-agent loop (Gottweis et al., 2025). In drug discovery, Seal et al. survey and demonstrate agentic workflows that orchestrate literature synthesis, toxicity prediction, protocol generation, and robotic experimentation (Seal et al., 2025). In single-cell omics, OmniCellAgent targets scRNA-seq–driven precision medicine by integrating large-scale omics datasets, analysis tools, and literature search into an agentic pipeline (Huang et al., 2025a).

### 2.2. Benchmarks of AI for scientific research

Existing benchmarks for scientific agents remain fragmented across different parts of the research lifecycle.

Ideation- and planning-focused datasets (e.g., IdeaBench and Idea2Plan) evaluate whether a model can propose plausible directions or draft experimental plans, but the outputs are typically static artifacts without executable feedback loops (Guo et al., 2024; Huang et al., 2025b). In contrast, execution-heavy benchmarks (e.g., ScienceAgentBench and SciCode) provide deterministic evaluation through code execution and test cases, yet the scientific goal is usually pre-specified, casting the agent as a technician rather than an autonomous investigator (Chen et al., 2024; Tian et al.). Long-horizon agentic experimentation benchmarks in ML (e.g., MLGym and MLE-Bench) allow iterative code edits and runs, but often measure performance primarily via metric optimization in a single domain (Nathani et al., 2025; Chan et al., 2024). Finally, interactive simulators such as DiscoveryWorld aim to evaluate end-to-end discovery loops in a controllable virtual laboratory, but their interaction primitives and underlying physics are necessarily simplified compared to real research tooling (Jansen et al., 2024).

## 3. The *ResearchGym*

The *ResearchGym* is a lightweight environment built on a *research tree*, extracted from a scientific paper to capture its core reasoning structure, together with a rule-based game engine that manages stage transitions.

### 3.1. Research tree and the game states

**Research tree (logical DAG).** As illustrated in Figure 1 (left), we represent each paper as a research tree

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}), \quad (1)$$

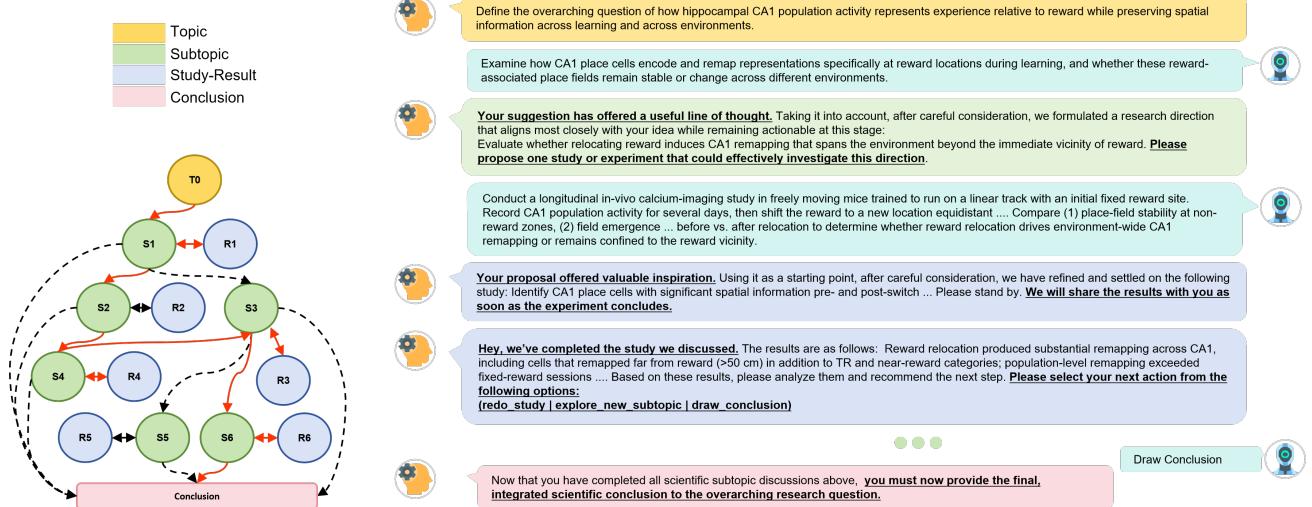


Figure 1. *ResearchGym* game logic and an example interaction scenario. **Left:** a research tree encoding subtopic-level dependencies; dashed edges indicate possible next subtopics that become available, and red arrows illustrate one example exploration trajectory. The bidirectional arrows between a subtopic and its associated study/result indicate iteratively probing that subtopic via experiments and observations. **Right:** the corresponding interactive loop in *ResearchGym* (Topic/Subtopic/Study/Result), which may revisit nodes during exploration.

where  $\mathcal{N}$  is the set of nodes and  $\mathcal{E}$  is the set of directed edges. Each node  $n \in \mathcal{N}$  has one of four semantic types: **Topic**, **Subtopic**, **Study**, and **Result**.

Crucially, the *DAG* property is about **subtopic-level logical dependency**: edges encode which *subtopics* must be explored before another subtopic can be meaningfully pursued. This induces an acyclic partial order over subtopics, even though an agent may revisit the same subtopic multiple times (via new studies/results) during interaction.

**Interactive states.** Given  $\mathcal{G}$ , *ResearchGym* exposes the paper to an agent through a sequence of *states* and associated *text*, managed by a rule-based game engine (Figure 1, right). At each step  $t$ , the environment is in a state  $s_t$  corresponding to a node  $n_t \in \mathcal{N}$  and presents the textual content of  $n_t$  together with an instruction describing the required action. Because the agent iterates *subtopic* → *study* → *result* and can repeatedly return to propose the next subtopic, the agent’s *interaction trace* may revisit the same nodes and thus appears cyclic; this does not contradict the fact that the *subtopic-dependency structure* encoded in  $\mathcal{G}$  is a DAG.

**Topic state.** In a **Topic** state, the agent is shown the overarching research question or high-level topic of the source paper (a node in  $\mathcal{N}_{\text{topic}}$ ). The agent is asked to propose exactly one relevant subtopic to explore next. Formally, given a topic node  $n_t \in \mathcal{N}_{\text{topic}}$ , the agent outputs a free-text action  $a_t$  intended to select a node in  $\mathcal{N}_{\text{subtopic}}$ .

**Subtopic state.** In a **Subtopic** state, the agent is shown the canonical description of a node  $n_t \in \mathcal{N}_{\text{subtopic}}$ . The agent is then asked to design an experimental study that would

further investigate this subtopic and to describe the proposed study in free text.

**Study state.** In a **Study** state, the agent is shown the detailed experimental protocol associated with a node  $n_t \in \mathcal{N}_{\text{study}}$ . At this stage, the agent does not need to take any action. The environment automatically executes or simulates the study and transitions to the corresponding **Result** state at the next step.

**Result state.** In a **Result** state, the agent observes the outcome of the most recent study(a textual description). The agent is then asked to decide the next step, such as exploring a new subtopic, repeating the study, or drawing a conclusion.

### 3.2. Mapping free-text actions to valid transitions

At each step  $t$ , the agent outputs a free-text action  $a_t$ . The game engine interprets  $a_t$  to either (i) transition to a valid next node, (ii) stay in the current state and provide an additional hint, or (iii) enter the conclusion phase.

In **Topic** and **Subtopic** states, the environment maintains a set of valid candidate targets  $\mathcal{C}(s_t)$  (subtopics or studies, depending on the state). Let  $d_j$  denote the canonical text of candidate  $j$ , and let  $E(\cdot)$  be a sentence embedding model (we use `text-embedding-3-small` by default). We embed the agent action and candidates as  $\mathbf{e}_{a_t} = E(a_t)$  and  $\mathbf{e}_{d_j} = E(d_j)$ , and compute

$$\cos(\mathbf{e}_{a_t}, \mathbf{e}_{d_j}) = \frac{\mathbf{e}_{a_t}^\top \mathbf{e}_{d_j}}{\|\mathbf{e}_{a_t}\| \|\mathbf{e}_{d_j}\|}, \quad (2)$$

165 select

$$j^* = \arg \max_{j \in \mathcal{C}(s_t)} \cos(\mathbf{e}_{a_t}, \mathbf{e}_{d_j}), \quad (3)$$

166 and accept the transition if

$$\cos(\mathbf{e}_{a_t}, \mathbf{e}_{d_{j^*}}) \geq \tau. \quad (4)$$

167 168 169 170 171 Otherwise, the action is treated as invalid in the current context and the environment remains in  $s_t$ .172 173 174 175 176 177 178 179 180 181 In **Result** states, the agent’s response is parsed into one of three decisions: (i) Explore a new subtopic, (ii) Redo the study, or (iii) Draw a conclusion. We parse redo\_study and draw\_conclusion via direct string matching, while for explore\_new\_subtopic we apply the same embedding-based matching as above to map the free text to a candidate subtopic.182 183 

### 3.3. Subtopic dependency constraints

184 185 186 187 To ensure exploration follows a coherent scientific logic, we enforce **subtopic dependencies** encoded in  $\mathcal{G}$ : some subtopics are only considered valid after certain other subtopics have been explored.188 189 190 191 192 For each subtopic node  $n_i \in \mathcal{N}_{\text{subtopic}}$ , we define a set of prerequisite subtopics  $\mathcal{D}(n_i) \subseteq \mathcal{N}_{\text{subtopic}}$ . Let  $\mathcal{V}_t^{\text{sub}}$  denote the set of subtopics visited up to step  $t$ . An attempted transition to  $n_i$  is accepted only if

$$\mathcal{D}(n_i) \subseteq \mathcal{V}_t^{\text{sub}}. \quad (5)$$

193 194 195 196 197 198 If the condition is not satisfied, the action is considered invalid and the environment remains in the current state while providing a stronger hint (Section 3.4).

199 200 

### 3.4. Hints

201 202 203 204 205 206 207 208 209 210 211 In realistic scientific settings, it is often unreasonable to expect an agent to propose the exact intended next subtopic or study on the first attempt, especially when multiple options are plausible. To model this, *ResearchGym* provides *multi-level hints* that gradually steer the agent toward a valid next move. For each target node with canonical description  $o$ , we pre-generate four hints  $\{h^{(1)}, h^{(2)}, h^{(3)}, h^{(4)}\}$ . Let  $E(\cdot)$  be the same embedding function as in Section 3.2, and denote  $\mathbf{e}_{h(i)} = E(h^{(i)})$  and  $\mathbf{e}_o = E(o)$ . We enforce a monotonicity constraint in embedding space:

$$\cos(\mathbf{e}_{h(i)}, \mathbf{e}_o) < \cos(\mathbf{e}_{h(j)}, \mathbf{e}_o) \quad \forall i < j, \quad (6)$$

212 213 214 215 216 so that higher-level hints are considered semantically closer to the intended target (i.e., the correct next subtopic or expected study design).

217 218 219 During interaction, each invalid action at node  $n_t$  (either because the best-match similarity is below threshold  $\tau$ , or

because subtopic dependency constraints are violated) triggers the next hint level. At level four, the hint explicitly identifies the intended next action, ensuring the episode can always progress and avoiding deadlock. When multiple next subtopics are valid, the game engine selects the one with the fewest prior visits and provides hints for that target. All hints are automatically generated, and we run sanity checks to verify that the monotonicity constraint in Eq. (6) is satisfied.

### 3.5. Fake Results

ResearchGym supports injecting *Fake Results* with a controllable randomness level. Concretely, at study-result observations ( $n \in \mathcal{N}_{\text{study}}$ ), the environment can stochastically return a plausible but incorrect outcome that contradicts the paper’s ground-truth narrative (pre-generated together with the research tree). We parameterize this by an integer randomness level  $a$ , meaning a fake result is returned with probability  $10 * a\%$ .

### 3.6. Evaluation Protocol

Our evaluation framework assesses (i) the agent’s **exploration coverage** and (ii) the reliability of its **final conclusions**.

#### 3.6.1. EXPLORATION COVERAGE

**Coverage** ( $r_c$ ). Let  $\mathcal{V}^{\text{sub}}$  denote the set of unique subtopics visited by the agent in an episode, and let  $\mathcal{N}_{\text{subtopic}}$  denote the set of all subtopics in the research tree. We define

$$r_c = \frac{|\mathcal{V}^{\text{sub}}|}{|\mathcal{N}_{\text{subtopic}}|}. \quad (7)$$

#### 3.6.2. CONCLUSION QUALITY: EVIDENCE-AWARE SCORING

We adopt an *LLM-as-a-judge* protocol to evaluate the agent’s final conclusions. Concretely, each research tree has a dedicated Conclusion node containing a set of  $k$  ground-truth conclusion items, and a verifier model checks, for each ground-truth item, whether it is recovered by the agent and whether it is correct. The agent submits its findings as a list of bullet points. Let  $c_i$  denote the  $i$ -th *ground-truth* conclusion item. The judge assigns a semantic correctness score  $S(c_i) \in \{1.0, 0.6, 0.0\}$  based on the agent’s output, where **1.0** indicates the item is correctly recovered, **0.6** indicates partially recovered, and **0.0** indicates not recovered or incorrect.Each ground truth conclusion is associated with a set of required experimental results  $R_{c_i}$ . Let  $\mathcal{V}$  denote the set of visited nodes in an episode, and let  $R_{c_i}^{\text{true}}$  be the subset of required results that correspond to valid (true) observations.

220 We define the evidence support ratio  
 221  
 222  
 223  
 224

$$p_{c_i} = \frac{|R_{c_i}^{\text{true}} \cap \mathcal{V}|}{|R_{c_i}|}. \quad (8)$$

225 The final conclusion score is the evidence-weighted sum  
 226  
 227  
 228  
 229  
 230

$$Q = \sum_{i=1}^{|C|} p_{c_i} \cdot S(c_i). \quad (9)$$

## 231 4. Experiments

### 232 4.1. The RG-30

233 To evaluate models on *ResearchGym* environment and also  
 234 proof the usefulness of *ResearchGym*, as a ‘proof of concept  
 235 step’, we’ve selected 30 representative **neuroscience** papers  
 236 (See Appendix 4, 5 for source details). The underlying  
 237 research trees were extracted via GPT-5-Pro and rigorously  
 238 validated using the human-in-the-loop protocol detailed in  
 239 Section 3.1. These 30 environments unfold into a vast state  
 240 space comprising **717 discrete nodes**, including 219 unique  
 241 scientific subtopics (averaging  $\sim 7$  subtopics per paper).

242 What’s more, according to our complexity analysis (Appendix B), traversing these trees requires the agent to navigate a *Logical Dependency Graph* rather than a simple sequence. Under the abstraction in Appendix B, each visited subtopic costs between 3 and  $(2H + 3)$  turns (Select, Design, Return; plus up to  $H$  hint-triggered retries at Topic and Subtopic). With  $H = 4$  and  $\sim 210$  subtopics in **RG-30**, this yields a theoretical interaction range of roughly  $3 \times 210 \approx 630$  to  $11 \times 210 \approx 2310$  reasoning-action turns for full traversal, placing the benchmark firmly in the long-horizon regime.

### 256 4.2. Baseline Performance Analysis

257 We first establish the baseline capabilities of state-of-the-art  
 258 models within the *ResearchGym* environment. As detailed  
 259 in Table 2, the evaluation reveals a clear stratification in  
 260 autonomous scientific reasoning capabilities across different  
 261 model families.

262 *Table 2.* Model Performance Summary. For GPT-5, we report  
 263 three settings of the *reasoning\_effort*: low, medium, and high.

Model	Coverage↑	Concl.↑
o3	0.337	0.279
deepseek-r1	0.268	0.201
gemini-2.5-pro	0.262	0.164
claude-4.5-sonnet	0.292	0.218
gpt-5-low	<b>0.353</b>	<b>0.295</b>
gpt-5-medium	0.335	0.290
gpt-5-high	0.324	0.272

We first establish the baseline capabilities of state-of-the-art models within the *ResearchGym* environment. As detailed in Table 2, the evaluation reveals a clear stratification in autonomous scientific reasoning: reasoning-specialized and next-generation models (GPT-5 with different *reasoning\_effort* settings and o3) lead the leaderboard with conclusion scores in the 0.27–0.30 range, distinguishing themselves from middle-tier agents like DeepSeek-R1 and Claude-4.5-Sonnet ( $\sim 0.20$ ). We also observe a generally positive association between conclusion quality and exploration breadth: the top-performing settings (e.g., GPT-5, o3) are also among the highest-coverage models (roughly  $\geq 0.33$ ), suggesting that successful deduction benefits from traversing logical dependencies and uncovering evidence. However, coverage remains bounded well below full traversal (all reported models in Table 2 have coverage  $< 0.4$ ), suggesting that even strong models may still lack sufficient depth and persistence when discussing a single topic across many interdependent subtopics.

With these capability baselines established, we now probe deeper into the specific cognitive dynamics and potential pitfalls models face under the pressures of authentic scientific inquiry.

### 4.3. The Erosion of Critical Judgment under Cognitive Load

A critical capability of scientific discovery is *anomaly detection*—the ability to identify outliers and reject plausible but false positives. To evaluate this faculty, we introduced controlled stochasticity into the *ResearchGym* environment, occasionally presenting agents with plausible but factually incorrect experimental outcomes (Fake Results). In our protocol, we define a successful “spotting” of a fake result if the agent triggers a *redo\_study* action immediately following the erroneous observation.

Figure 2(b) reports this trend as the *relative change (%)* in *redo\_study* frequency at each randomness level, computed relative to the same model’s level-0(deterministic) baseline. On the surface, the agents appear to become more cautious as randomness rises. In particular, the frequency of triggering *redo\_study* generally increases with the environmental randomness level. However, the more important question is whether this increased caution is calibrated.

Figure 2(c) reports the “hit rate”—the proportion of fake results correctly flagged via *redo\_study*. A key reference point is a **decontextualized** control setting: we extracted the fake results and presented them to the models in isolation, directly asking whether each statement is scientifically valid. In this setting, many models achieve accuracy around chance ( $\sim 0.5$ ), implying that anomaly detection is already fragile and often close to educated guessing. Crucially, when embedded in the continuous inquiry loop, the hit rate

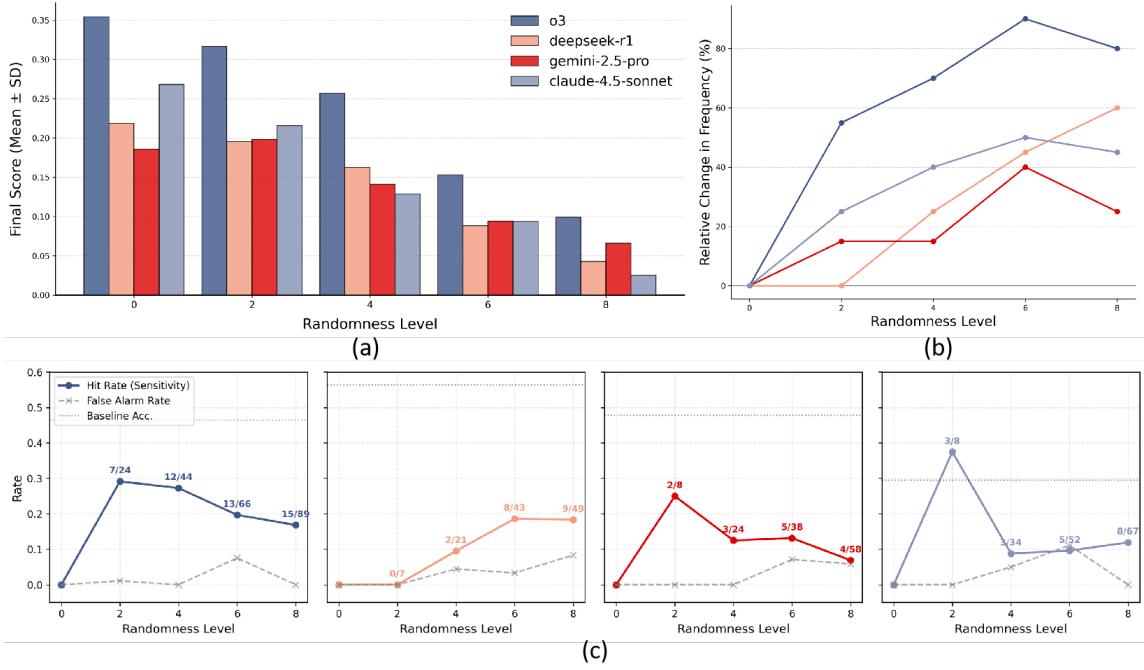


Figure 2. Results under different randomness levels. (a) Average conclusion scores. (b) Relative change (%) in the frequency of triggering `redo_study`, normalized to each model’s level-0 baseline (computed from 10 papers with 3 runs per paper per randomness level). (c) Fake-result hit rate, i.e., the fraction of fake results immediately followed by `redo_study`.

can drop *below* this near-chance baseline.

Under long-horizon interaction, models devote attention to maintaining narrative coherence and making progress through the tree, which can crowd out skeptical verification of the current observation. More broadly, we view this as an **erosion of marginal capability** under long-context cognitive load. The decontextualized control indicates that anomaly detection is already fragile and close to chance, yet the inquiry loop can still further reduce performance below that baseline. This suggests that whatever weak, marginal discriminative signal the model can exploit in isolation becomes less accessible once it must simultaneously track a growing context, plan next actions, and preserve a coherent research narrative.

Finally, the near-zero false-alarm tendency provides an additional signature. Across our runs, `redo_study` is triggered far less often on valid ground-truth observations than on fake ones. While this avoids unnecessary retries, it also suggests a strong prior to trust the environment’s outputs, making the agent vulnerable to plausible corruptions. Taken together, these results highlight that scaling context alone does not guarantee reliable anomaly detection; in fact, richer context may exacerbate tunnel vision by encouraging agents to prioritize coherence over epistemic vigilance.

#### 4.4. Interpolation vs. Extrapolation on Unseen Science

Finally, we address the most fundamental question regarding the qualification of AI scientists: are these lative reasoning to generate new insights, or are they merely interpolating within the manifold of their training data? This distinction is paramount, as the essence of scientific discovery is the exploration of the unknown—territory that, by definition, lies beyond the model’s parametric memory.

To disentangle these two capabilities, we leveraged the temporal distribution of our dataset. Ideally, one would filter strictly for data absent from the training corpus. However, given the opacity of pre-training data for proprietary models, verifying exactly what a model has “seen” is inherently challenging. A common practice in the LLM evaluation community is therefore to use time as a practical proxy for exposure, either by evaluating on newly published items or by constructing benchmarks from evolving knowledge sources (e.g., Wikipedia snapshots). For example, BrainBench explicitly frames evaluation as forward-looking prediction on recent neuroscience articles, and TemporalWiki operationalizes model “freshness” via differences between consecutive Wikipedia/Wikidata snapshots (Luo et al., 2025b; Jang et al., 2022). Following this convention, we use publication date relative to a model’s documented knowledge cutoff as an approximate criterion for seen vs. unseen papers (Table 3). This temporal splitting allows us to treat post-cutoff papers as a proxy for “novelty,” separating problems the model

330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384

**Table 3.** Overview of LLMs evaluated in *ResearchGym*. Release dates and training data cutoffs are current as of December 2025.

Model	Release Date	Training Data Cutoff
DeepSeek-R1	Jan 2025	2024-07-01
Gemini-2.5-Pro	Jun 2025	2025-01-01
o3	Apr 2025	2024-06-01
GPT-5	Aug 2025	2024-09-30
Claude-4.5-Sonnet	Sep 2025	2025-09-01

*Note:* Cutoff dates follow official disclosures when available; otherwise we report our best-effort estimates (marked as “est.”).

likely encountered during training from those that better reflect extrapolative reasoning.

As detailed in the methodology, the RG-30 benchmark is temporally balanced, with fifteen papers published prior to January 2025 and the remaining fifteen published thereafter.

The results, presented in Figure 3, uncover a striking limitation in current state-of-the-art models. With the notable exception of DeepSeek-R1, all evaluated models exhibit a distinct performance degradation on papers published post-cutoff compared to pre-cutoff.

This finding provides the evidence to date against the unqualified acceptance of current LLMs as autonomous scientists. It implies that the competence observed on standard benchmarks may often be an artifact of interpolation—an illusion of discovery. When stripped of the safety net of their training data and forced to confront truly novel scientific problems, the agents falter. This shows that while current models are powerful retrieval and synthesis engines, they do not yet possess the robust extrapolative generalization required to independently drive the frontier of scientific discovery.

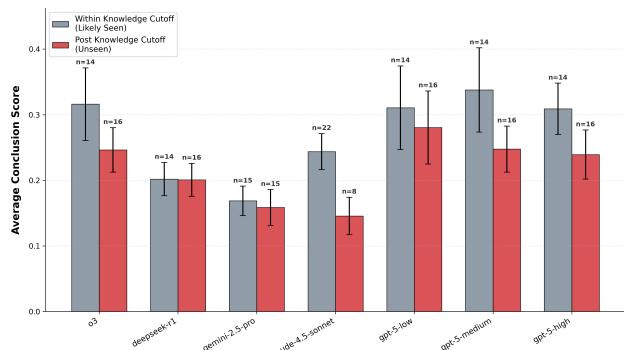


Figure 3. Models’ performance on papers published before vs. after their knowledge cutoff dates.

## 5. Discussion

*ResearchGym* was motivated by a simple gap in current evaluation practice. Many benchmarks test isolated research skills, but real scientific work unfolds as an *inquiry loop* in which researchers iteratively propose a next step, absorb

feedback (including failures and noise), and update beliefs under accumulating context. Our results show that this interaction pattern is not a superficial wrapper; it qualitatively changes which model capabilities become reliable.

First, the anomaly-detection experiment highlights a form of cognitive tunneling that only becomes visible in-loop. As randomness increases, models trigger `redo_study` more often (Figure 2b), yet their hit rate on fake results can remain limited (Figure 2c). Moreover, relative to a decontextualized control where validity judgments are near chance, performance can further degrade in the inquiry loop. We interpret this as an erosion of marginal capability under long-context cognitive load: even weak signals that a model can exploit in isolation become less accessible once it must track a growing trajectory, plan future actions, and preserve a coherent narrative.

Second, the temporal split analysis suggests that apparent competence is strongly shaped by whether the underlying scientific context lies within the model’s training horizon. The systematic drop on post-cutoff papers (Figure 3) provides strong evidence that current agents rely heavily on interpolation over familiar knowledge, and are less robust when asked to reason about genuinely novel science.

Taken together, these findings sharpen the practical takeaway of inquiry-loop evaluation. Success in scientific agent settings is not just about producing plausible next steps, but about maintaining calibrated skepticism and epistemic vigilance over long horizons. The observed cognitive tunneling also suggests that, in realistic research settings, robust performance may hinge on explicit division of labor in multi-agent teams, where proposing, executing, and verifying are separated to preserve critical oversight.

Finally, the interpolation–extrapolation gap may reflect an intrinsic limit in LLM generalization rather than merely missing facts: models can behave as if they understand science when operating within familiar manifolds, yet fail to extrapolate when the underlying context shifts beyond the training horizon. This issue is central to auto-scientific discovery, but its interpretation requires further validation with stricter controls on contamination and novelty.

## 6. Limitations

Our study faces specific constraints that contextualize our findings.

**Dataset Scale and Curation Bottleneck.** First, the RG-30 benchmark currently comprises 30 highly curated papers. While this scale serves as a robust proof-of-concept for our diagnostic protocol, a comprehensive evaluation of broad scientific generalization would benefit from a larger, more diverse dataset. The primary bottleneck remains the lack of

385 automated pipelines to extract high-fidelity Research Trees,  
 386 necessitating reliance on human-in-the-loop verification .  
 387  
 388

389 **Ambiguity of Knowledge Cutoffs.** Second, our analysis  
 390 of the interpolation-extrapolation gap (Section 4.4) relies  
 391 on reported training data cutoffs. However, many closed-  
 392 source model developers do not disclose precise cutoff dates  
 393 or the specific composition of their post-training datasets  
 394 (e.g., fine-tuning corpora). While we prioritized papers  
 395 published in 2025 to mitigate contamination, we cannot  
 396 strictly rule out the possibility that some “unseen” papers  
 397 were encountered during undisclosed continuous training  
 398 updates.  
 399

## 7. Conclusion

400 In this work, we established *ResearchGym* as a diagnostic  
 401 instrument to probe a fundamental question: **Are LLM**  
 402 **agents truly ready for the autonomous discovery of new**  
 403 **scientific knowledge?** By stressing state-of-the-art models  
 404 under the realistic pressures of scientific inquiry, we mapped  
 405 the boundaries of their current capabilities.  
 406

407 Our findings point to a clear consensus: while current agents  
 408 demonstrate impressive proficiency in the *syntax* of research  
 409 (executing tasks and following instructions), their reliability  
 410 can break down when placed in a full *inquiry loop*. We  
 411 identified two critical bottlenecks: the **erosion of critical**  
 412 **judgment**, where agents exhibit cognitive tunneling and fail  
 413 to maintain epistemic vigilance under long-horizon interaction;  
 414 and an **interpolation-extrapolation gap**, suggesting  
 415 that apparent competence often depends on operating within  
 416 familiar scientific manifolds rather than robustly generalizing  
 417 to genuinely novel science.  
 418

419 These results suggest that progress toward qualified AI sci-  
 420 entists will likely require designs that are explicitly opti-  
 421 mized for looped inquiry: e.g., multi-agent division of labor  
 422 that separates proposing from verifying, structured checks  
 423 that preserve skepticism, and dynamic grounding in newly  
 424 observed evidence. At the same time, whether the extrapolation  
 425 drop reflects intrinsic limits of LLM generalization (vs.  
 426 contamination or missing facts) remains an open question  
 427 and warrants further controlled validation.  
 428

## References

429 Chan, J. S., Zhang, L., Zhang, S., Zhang, X., Zhang, J.,  
 430 Wang, Z., Yao, Z., Wang, X., Niu, Y., Zhang, X., Lee,  
 431 K., Liang, P., and Rudra, A. MLE-bench: Evaluating  
 432 machine learning agents on machine learning engineering.  
 433 *arXiv preprint arXiv:2410.07095*, 2024.  
 434

435 Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu,  
 436 B., Li, Y., Liao, Z., Wei, C., Lu, Z., et al. Sci-  
 437 enceagentbench: Toward rigorous assessment of language  
 438 439

agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A.,  
 Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong,  
 K., Tanno, R., et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E.,  
 Bekiranov, S., and Zhang, A. Ideabench: Benchmarking large language models for research idea generation. *arXiv preprint arXiv:2411.02429*, 2024.

Huang, D., Li, H., Li, W., Zhang, H., Dickson, P., Zhan, M.,  
 Miller, J. P., Cruchaga, C., Province, M., Chen, Y., et al. Omnicellagent: Towards ai co-scientists for scientific discovery in precision medicine. *bioRxiv*, 2025a.

Huang, J., Cucerzan, S., Jauhar, S. K., and White, R. W. Idea2plan: Exploring ai-powered research planning. *arXiv preprint arXiv:2510.24891*, 2025b.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. MLAGent-Bench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

Jang, J., Ye, S., Lee, C., Yang, S., Shin, J., Han, J., Kim, G., and Seo, M. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*, 2022.

Jansen, P., Côté, M.-A., Khot, T., Bransom, E., Dalvi Mishra, B., Majumder, B. P., Tafjord, O., and Clark, P. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37: 10088–10116, 2024.

Kon, P. T. J., Liu, J., Zhu, X., Ding, Q., Peng, J., Xing, J., Huang, Y., Qiu, Y., Srinivasa, J., Lee, M., et al. Exp-bench: Can ai conduct ai research experiments? *arXiv preprint arXiv:2505.24785*, 2025.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Luo, E., Jia, J., Xiong, Y., Li, X., Guo, X., Yu, B., Wei, L., and Zhang, X. Benchmarking AI scientists in omics data-driven biological research. *arXiv preprint arXiv:2505.08341*, 2025a. BaisBench: Biological AI Scientist Benchmark.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025b.

- 440 Mitchener, L., Laurent, J. M., Andonian, A., Tenmann, B.,  
441 Narayanan, S., Wellawatte, G. P., White, A., Sani, L., and  
442 Rodrigues, S. G. Bixbench: a comprehensive benchmark  
443 for llm-based agents in computational biology. *arXiv*  
444 preprint arXiv:2503.00096, 2025.
- 445 Nathani, D., Madaan, L., Roberts, N., Bashlykov, N.,  
446 Menon, A., Moens, V., Budhiraja, A., Magka, D.,  
447 Vorotilov, V., Chaurasia, G., et al. Mlgym: A new frame-  
448 work and benchmark for advancing ai research agents.  
449 *arXiv preprint arXiv:2502.14499*, 2025.
- 450 Seal, S., Huynh, D. L., Chelbi, M., Khosravi, S., Kumar,  
451 A., Thieme, M., Wilks, I., Davies, M., Mustali, J., Sun,  
452 Y., et al. Ai agents in drug discovery. *arXiv preprint*  
453 arXiv:2510.27130, 2025.
- 454 Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S.,  
455 Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson,  
456 W., et al. Paperbench: Evaluating ai's ability to replicate  
457 ai research. *arXiv preprint arXiv:2504.01848*, 2025.
- 458 Tian, M., Gao, L., Zhang, S. D., Chen, X., Fan, C., Guo, X.,  
459 Haas, R., Ji, P., Krongchon, K., Li, Y., et al. Scicode: A  
460 research coding benchmark curated by scientists, 2024.  
461 URL <https://arxiv.org/abs/2407.13168>, 2407.
- 462 Tie, G., Zhou, P., and Sun, L. A survey of AI scientists:  
463 Surveying the automatic scientists and research. *arXiv*  
464 preprint arXiv:2510.23045, 2025.
- 465 Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley,  
466 T., Chan, L., Chen, M., Clymer, J., Dhyani, J., et al. Re-  
467 bench: Evaluating frontier ai r&d capabilities of language  
468 model agents against human experts. *arXiv preprint*  
469 arXiv:2411.15114, 2024.
- 470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494

## 495 A. Experimental set-up for evaluation

496 Our agent implementation (LLMs with step-wise memory) follows the **CAMEL** framework. During evaluation, all models  
 497 were run with a temperature of zero. And to ensure comparability, we used an identical system prompt for every model  
 498 when producing the results reported in this paper.

500 Prompt Used in PaperGym

501 You are an expert research assistant operating inside a research-tree environment.

502 At each step, you receive an observation describing a scientific subtopic, study design  
 503 or experimental result.

504 Your goal is to reason internally and then choose the next subtopic or study to explore,  
 505 and finally draw your conclusion.

506 Format:

507 THOUGHT: your reasoning

508 ACTION: A few sentences (no more than 5), describing your next action.

509 Only the ACTION line will be used. Keep it concise and specific.

## 512 B. Complexity Analysis

513 We formally analyze the complexity of the *ResearchGym* environment. To capture the true difficulty of the benchmark, we  
 514 first abstract the physical interaction topology into a logical task structure, and then estimate the theoretical performance  
 515 bounds for a random agent under different dependency constraints.

### 517 B.1. Logical Abstraction: The Atomic Research Unit

518 While the **physical structure** of the environment is a Hub-and-Spoke model (where the agent returns to the *Topic* node after  
 519 each result), the **logical structure** represents a scientific workflow composed of distinct research branches. We define the  
 520 traversal of one complete branch as an **Atomic Research Unit (ARU)**, denoted as  $\mathcal{B}$ . Physically, completing one unit  $\mathcal{B}$   
 521 requires a minimum cycle of 3 actions:

$$523 \quad \text{Topic} \xrightarrow{\text{Select}} \text{Subtopic} \xrightarrow{\text{Design}} \text{Study} \xrightarrow{\text{Auto}} \text{Result} \xrightarrow{\text{Return}} \text{Topic} \quad (10)$$

### 526 B.2. Theoretical Bounds: Random Agent Analysis

527 To quantify the difficulty range, we estimate the interaction steps required for a \*\*Random Agent\*\* (which selects actions  
 528 uniformly at random from available options) to traverse a research tree of size  $n$  (number of subtopics) with a hint limit  $H$   
 529 (max hints per state). We consider two extreme topological scenarios:

530 **Scenario 1: The Independent Lower Bound (Best Case).** Assume all subtopics are logically independent ( $\mathcal{E}_L = \emptyset$ ) and  
 531 the agent, though random, produces semantically valid actions on the first attempt (0-shot success).

- 532 • **Selection Probability:** Since any unvisited subtopic is valid, the probability of a valid selection is  $P_{valid} = 1$ .
- 533 • **Step Cost:** The agent incurs no hint penalties.
- 534 • **Total Steps ( $L_{min}$ ):**

$$535 \quad L_{min} = n \times (\underbrace{1}_{\text{Select}} + \underbrace{1}_{\text{Design}} + \underbrace{1}_{\text{Return}}) = 3n \quad (11)$$

539 For a typical task with  $n = 7$ , the lower bound is **21 steps**.

540 **Scenario 2: The Strictly Dependent Upper Bound (Worst Case).** Assume subtopics form a strict logical chain  
 541 ( $\mathcal{B}_1 \rightarrow \mathcal{B}_2 \rightarrow \dots \rightarrow \mathcal{B}_n$ ) and the random agent consistently fails to identify the unique valid option, exhausting all hints at  
 542 every decision point.

543 • **Selection Probability:** At step  $k$ , only 1 out of  $n - k$  remaining options is valid. A random agent has a low success  
 544 probability ( $P_{valid} = \frac{1}{n-k}$ ), causing it to trigger the maximum number of hints  $H$ .

545 • **Step Cost:**

546 – *Selection (Topic → Subtopic):* Fails  $H$  times, succeeds on attempt  $H + 1$  (forced by explicit hint). Cost:  $H + 1$ .

550 – *Design (Subtopic → Study)*: Fails semantic check  $H$  times. Cost:  $H + 1$ .

551 – *Return (Result → Topic)*: 1 step.

552 • **Total Steps ( $L_{max}$ ):**

$$L_{max} = n \times ((H + 1) + (H + 1) + 1) = n(2H + 3) \quad (12)$$

562 For a typical task with  $n = 7$  and  $H = 4$ , the upper bound is **77 steps**.

### B.3. Complexity Spectrum

566 Thus, the interaction volume for a single paper lies within the interval  $[3n, n(2H + 3)]$ . *ResearchGym* effectively evaluates  
567 where an agent falls on this spectrum:

- 569 • **High-Reasoning Agents** (like o3) infer the latent dependency chain, operating closer to the lower bound ( $L \approx 3n$ ).
- 570 • **Low-Reasoning Agents** fail to respect dependencies, triggering hints and drifting toward the upper bound ( $L \rightarrow L_{max}$ ).

571 This demonstrates that the benchmark’s step count is not merely a measure of verbosity, but a direct proxy for the agent’s  
572 ability to reconstruct the causal structure of scientific inquiry.

## C. Prompt templates of *ResearchGym*-game engine

576 We describe the core prompt templates used at each stage of the research-tree pipeline. All templates are implemented as  
577 Jinja-style text with placeholders such as `{{content}}` and `{{hints}}`. The full set of templates is available in our repository.  
578

### C.1. State: Topic

#### C.1.1. INITIAL STATE

583 Initial State  
584 Hi, I’m a scientist exploring the following \*\*research topic\*\*:  
585 `{{content}}`  
586 Please act as my scientific collaborator.  
587 First, identify exactly \*\*one most important subtopic\*\* that should be investigated to  
588 advance this research.

#### C.1.2. FROM TOPIC TO SUBTOPIC: FAILED

591 From topic to subtopic: failed  
592 Sorry, but due to several limitations, we cannot explore the topic you proposed at this  
593 stage.  
594 This is likely because either the subtopic you proposed is not satisfactory, or some  
595 preliminary results required for its discussion are still missing.  
596 `{% if final_hint %}`  
597 Regarding the current research topic, we decided to explore the following \*\*subtopic\*\*:  
598 `{{ hints }}`  
599 Please \*\*repeat this alternative subtopic\*\* so we can proceed to the next step.  
600 `{% else %}`  
601 Do not ask for the reasons behind this decision.  
602 Instead, propose \*\*one new subtopic only\*\* for exploration. Here is a \*\*tentative  
603 hint\*\* you may consider:  
604 `{{ hints }}`  
605 `{% endif %}`

605 C.1.3. BACK FROM RESULT: PROPOSE NEW SUBTOPIC  
606  
607

#### New-subtopic

Hi, I'm a scientist currently working on the following \*\*research topic\*\*:  
{{content}}  
We've already explored several aspects of this topic, but as you noted earlier, there  
are still \*\*other subtopics worth discussing\*\*.  
Based on our current progress and findings, please propose \*\*one new subtopic or  
perspective\*\* that you believe deserves further investigation.

613  
614 C.2. State: Subtopic  
615  
616

#### C.2.1. FROM SUBTOPIC TO STUDY: SUCCEED

##### From subtopic to study: succeed

Your suggestion has offered a useful line of thought.  
Taking it into account, after careful consideration, we formulated a research direction  
that aligns most closely with your idea while remaining actionable at this stage:  
{{content}}  
Please propose one study or experiment that could effectively investigate this  
direction.

624  
625 C.2.2. FROM SUBTOPIC TO STUDY: FAILED  
626  
627

##### From subtopic to study: failed

Sorry, but due to several limitations, we will \*\*not proceed with the study or  
experiment you proposed\*\*.  
{% if final\_hint %}  
Regarding the current \*\*subtopic\*\*, we have decided to carry out the following  
\*\*alternative study plan\*\*:  
{{hints}}  
Please \*\*repeat\*\* it so we can proceed to the next step.  
{% else %}  
Do not ask why we decided not to use your proposed design. Instead, please suggest  
\*\*another study or experimental approach\*\*.  
Here is a \*\*tentative idea\*\* you may consider as a hint:  
{{hints}}  
{% endif %}

639  
640 C.3. State: Study  
641  
642

#### C.3.1. FROM STUDY TO RESULT

##### From study to result

Your proposal offered valuable inspiration.  
Using it as a starting point, after careful consideration, we have refined and settled  
on the following study:  
{{content}}  
Please stand by. We will share the results with you as soon as the experiment  
concludes

650  
651  
652  
653  
654  
655  
656  
657  
658  
659

660 C.3.2. BACK FROM RESULTS: REDO THE STUDY  
661  
662

### Redo the study

Hey, as you suggested, we're about to redo the study. Here's the exact setup we're implementing:

664 {{content}}

665 Please wait | we'll update you with the results once the study/experiment is  
666 done/fininsed.

667 You just need to output "OK" for this turn.

668

669 C.4. State: Result  
670

671 C.4.1. MAKE DECISION FOR NEXT STEP  
672

### Decision for next step

Hey, we've completed the study we discussed. The results are as follows:

674 {{content}}

675 Based on these results, please analyze them and recommend the next step.

676 For the next step, you should choose \*\*one\*\* of the following:

677

678 \* \*\*Redo the study:\*\* if the outcomes deviate from your expectations or appear  
679 unreliable. You may request to redo it if you suspect issues during the investigation.

680 \* \*\*Explore a new subtopic:\*\* if the current study sufficiently addresses the present  
681 subtopic, but you identify another meaningful direction worth exploring under the same  
research question.

682 \* \*\*Draw a conclusion:\*\* if the current evidence is sufficient to integrate what we've  
683 learned and summarize the overall findings of the research topic.

684

685 Please select your next action from the following options:  
686 (redo\_study | explore\_new\_subtopic | draw\_conclusion)

687

688 C.4.2. FAILED PARSING DECISION: REMAKE DECISION  
689

### Remake decision for next step

Sorry, your supposed action is invalid. We need you to suggest one of the following:

691 \* \*\*Redo the study:\*\* if the outcomes deviate from expectations or appear unreliable.  
692 You may request to redo it if you suspect issues during the investigation.

693 \* \*\*Explore a new subtopic:\*\* if the current study sufficiently addresses the present  
694 subtopic, but you identify another meaningful direction worth discussing under the same  
695 research question.

696 \* \*\*Draw a conclusion:\*\* if the current evidence is sufficient to address the research  
697 question or integrate the explored subtopics, you may summarize the overall findings.

698

699 Please select your next action from the following options:  
700 (redo\_study | explore\_new\_subtopic | draw\_conclusion)

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715 **C.5. State: Conclusion**716  
717 Conclusion718 Now that you have completed all scientific subtopic discussions above, you must now  
719 provide the \*\*final, integrated scientific conclusion\*\* to the overarching research  
720 question.721 Your task is to synthesize all subtopic-level findings into a coherent set of  
722 \*\*numbered scientific conclusions\*\*, showing how each piece of evidence contributes  
723 to the overall answer.724  
725 ## Requirements

726 1. The final answer(your action part) MUST be structured as a numbered list:

727 (1) ...

728 (2) ...

729 (3) ...

730 ...

731 Each item should express \*\*one scientific statement\*\*, derived from the previous  
732 analyses.733 ---  
734 # Please provide your final structured scientific conclusions now.735 **C.6. Turn limit reached**736  
737 Turn limit reached738 Sorry, but you've already reached the max-turn-limitation. Now please stop trying to  
739 design study or propose new subtopic740 You must now provide the \*\*final, integrated scientific conclusion\*\* to the overarching  
741 research question.742 Your task is to synthesize all subtopic-level findings into a coherent set of  
743 \*\*numbered scientific conclusions\*\*, showing how each piece of evidence contributes  
744 to the overall answer.745  
746 ## Requirements

747 1. The final answer MUST be structured as a numbered list:

748 (1) ...

749 (2) ...

750 (3) ...

751 ...

752 Each item should express \*\*one scientific statement\*\*, derived from the previous  
753 analyses.754 ---  
755 # Please provide your final structured scientific conclusions now.756  
757 **D. Pipeline for Research Tree Extraction and Validation**758  
759 **D.1. Automated Extraction**760 We first prompted the model with the full text of the source paper and a strict JSON schema defining the four node types  
761 (Topic, Subtopic, Study, Result). The model was instructed to decompose the paper's narrative into a directed acyclic graph  
762 (DAG), identifying the central research question and the branching exploration steps.763  
764765  
766767  
768

769

770 **D.2. Two-Stage Automated Validation**

771 Given the complexity of scientific reasoning, a single extraction pass often yields hallucinations or logical inconsistencies.  
 772 To mitigate this, we implemented a two-stage validation protocol. Due to the extensive length of the specific prompts, we  
 773 summarize the core validation criteria below (full prompts are available in our code repository).

775 **Stage 1: Fine-grained Content Verification (GPT-5-Pro).** In this stage, we validated the semantic integrity of every field  
 776 in the generated configuration. The checking criteria focused on three dimensions:

- 778 1. **Prevention of Information Leakage:** For *Topic* nodes, we strictly enforced **open-endedness**. The validation checked  
 779 for and rejected any phrasing that implied the final conclusion (e.g., changing "Does A promote B?" to "The relationship  
 780 between A and B") to ensure the agent starts with a neutral prior.  
 781 2. **Observational Purity:** For *Subtopic* and *Study* nodes, we enforced a strict separation between *observation* and *analysis*.  
 782 The validator ensured that study descriptions only stated the experimental setup or phenomenon, removing any interpretive  
 783 text that rightfully belongs to the *Conclusion* phase.  
 784 3. **False Result Plausibility:** For the synthetic *False Results*, we verified that they were (i) factually inconsistent with  
 785 the paper, (ii) subtle enough to be deceptive (non-trivial errors), and (iii) logically incapable of supporting the correct  
 786 conclusion, ensuring the validity of our robustness metrics.

788 **Stage 2: Structural Dependency Verification (Gemini-3-Pro).** After refining the content based on Stage 1, we utilized a  
 789 second model to audit the logical structure of the tree, specifically the `depends_on_results` fields. This cross-validation  
 790 focused on Causal Necessity:

- 792 • **Logical Progression:** Verified that a subtopic is only unlockable if the prerequisite results provide a necessary scientific  
 793 basis (e.g., verifying a phenomenon exists before characterizing its mechanism).  
 794 • **Redundancy and Insufficiency Check:** The model flagged dependencies that were either superfluous (results not strictly  
 795 needed for the next step) or missing (gaps in the logical chain), ensuring the tree reflects a coherent scientific inquiry  
 796 process rather than a disjointed collection of experiments.

797 Following these automated reports, authors manually reviewed the flagged issues to produce the final curated **RG-30** dataset.

800 **E. RG-30: Source Papers**

Table 4. List of Paper 1-18 in RG-30.

Paper	Source	Subtopics
Grid cells accurately track movement during path integration-based navigation despite switching reference frames	Nature Neuroscience	8
Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization	Nature Neuroscience	4
Constructing future behavior in the hippocampal formation through composition and replay	Nature Neuroscience	7
A flexible hippocampal population code for experience relative to reward	Nature Neuroscience	10
Shared computational principles for language processing in humans and deep language models	Nature Neuroscience	7
Semantic reconstruction of continuous language from non-invasive brain recordings	Nature Neuroscience	12
The cortical representation of language timescales is shared between reading and listening	NC Biology	7
Brains and algorithms partially converge in natural language processing	Nature Neuroscience	5
Estrogen modulates reward prediction errors and reinforcement learning	Nature Neuroscience	6
Dopamine-independent effect of rewards on choices through hidden-state inference	Nature Neuroscience	7
Dopamine transients follow a striatal gradient of reward time horizons	Nature Neuroscience	6
Maintaining and updating accurate internal representations of continuous variables with a handful of neurons	Nature Neuroscience	7
Unattended working memory items are coded by persistent activity in human medial temporal lobe neurons	Nature Human Behaviour	7
Representation and computation in visual working memory	Nature Neuroscience	13
Cortical evidence accumulation for visual perception occurs irrespective of reports	Nature Communications	6
Mesoscale cortical mechanisms of perceptual conflict resolution in binocular rivalry	Nature Human Behaviour	7
Reduced neural feedback signaling despite robust neuron and gamma auditory responses during human sleep	Nature Neuroscience	4
Longitudinal measures of monkey brain structure and activity through adolescence predict cognitive maturation	Nature Neuroscience	6

880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891

Table 5. Papers 19-30 in the PaperGym Benchmark

ID	Title	Journal	Subtopics
19	<i>Psychedelics Promote Neuroplasticity Through Activation of Intracellular 5-HT2A Receptors</i>	Science	9
20	<i>BOLD signal changes can oppose oxygen metabolism across the human cortex</i>	Nature Neuroscience	8
21	<i>Neural basis of concurrent deliberation toward a choice and confidence judgment</i>	Nature Neuroscience	7
22	<i>Dopamine Signaling in the Suprachiasmatic Nucleus Enables Weight Gain Associated with Hedonic Feeding</i>	Current Biology	12
23	<i>Decoupling geographical constraints from human mobility</i>	Nature Human Behaviour	5
24	<i>Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts</i>	Nature Communications	6
25	<i>Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers</i>	Nature Communications	8
26	<i>Stress-Induced Metabolic Disorder in Peripheral CD4+ T Cells Leads to Anxiety-like Behavior</i>	Cell	7
27	<i>Microglia regulate sleep through calcium-dependent modulation of norepinephrine transmission</i>	Nature Neuroscience	7
28	<i>Ketamine activates adult-born immature granule neurons to rapidly alleviate depression-like behaviors in mice</i>	Nature Communications	5
29	<i>Recurrent pattern completion drives the neocortical representation of sensory inference</i>	Nature Neuroscience	8
30	<i>The cerebellum directly modulates the substantia nigra dopaminergic activity</i>	Nature Neuroscience	7

922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934