InquiTree: Evaluating AI Agents in the Scientific Inquiry Loop with Paper-Derived Research Trees

Anonymous authors
Paper under double-blind review

Abstract

While LLM-based agents are increasingly used in scientific workflows, it remains unclear whether they are truly qualified for the dynamic and uncertain process of discovery. Existing static evaluations often conflate genuine reasoning with rote memorization. We introduce *InquiTree*, a diagnostic environment that formalizes scientific inquiry as interactive Research Trees: directed acyclic graphs capturing the logical dependencies among hypothesis formulation, study design, result interpretation, and belief updating. Evaluating agents on a 30-paper test pool and releasing the open-access IT-18 subset, we identify two key limitations. First, agents exhibit an "Erosion of Marginal Capabilities": during long-horizon interactions, they develop "cognitive tunneling," where critical judgment and anomaly detection degrade relative to their intrinsic baselines. Second, performance drops on papers published after model training cutoffs, revealing a boundary between interpolation and extrapolation and suggesting that apparent competence is partly driven by parametric memory. These findings indicate that scaling context alone is insufficient for reliable AI scientists; stronger architectures or human oversight may be required to preserve critical evaluation and generalization.

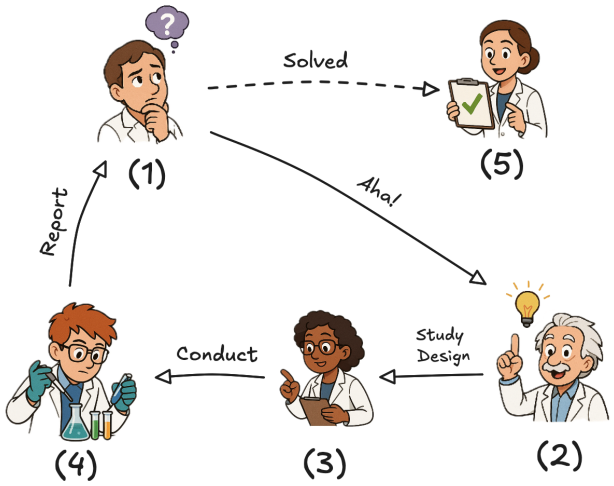

Figure 1: *InquiTree* abstracts scientific discovery as an inquiry loop over paper-derived research trees, where agents iteratively propose directions, receive study feedback, and update conclusions.

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*,

Benchmark	Subtopic Proposal	Study Design	Study Execution	Result Analysis
IdeaBenchGuo et al. (2024)	✓	✗	✗	✗
Idea2PlanHuang et al. (2025b)	✗	✓	✗	✗
ScienceAgentBenchChen et al. (2024)	✗	✗	✓	✗
SciCodeTian et al. (2024)	✗	✗	✓	✗
BAISBenchLuo et al. (2025a)	✗	✗	✓	✓
MLAgentBenchHuang et al. (2023)	✗	✗	✓	✓
MLGym-BenchNathani et al. (2025)	✗	✗	✓	✓
MLE-BenchChan et al. (2024)	✗	✗	✓	✓
BixBenchMitchener et al. (2025)	✗	✗	✓	✓
RE-BenchWijk et al. (2024)	✗	✗	✓	✓
EXP-BenchKon et al. (2025)	✗	✗	✓	✓
PaperBenchStarace et al. (2025)	✗	✗	✓	✓
DiscoveryWorldJansen et al. (2024)	✓	✓	✓	✓
IT-18 (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.

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. (2024). 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 *InquiTree*, an inquiry-loop evaluation environment that models research as an interactive process rather than a single-shot task. *InquiTree* 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 for evaluation and release **IT-18**, an 18-paper open-access subset whose configurations and logs can be publicly distributed. For non-open-access papers, we report aggregate test results but do not release paper-derived configurations or logs.

Contributions. We make the following contributions:

- We introduce *InquiTree*, 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 a 30-paper evaluation pool and release **IT-18**, an 18-paper open-access benchmark subset with configurations and logs; for non-open-access papers, we report test results without releasing paper-derived configs or logs.
- 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. (2024). 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 *InquiTree*

The *InquiTree* 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 2 (left), we represent each paper as a research tree

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

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 $\mathcal{G}$, *InquiTree* exposes the paper to an agent through a sequence of *states* and associated *text*, managed by a rule-based game engine (Figure 2, 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

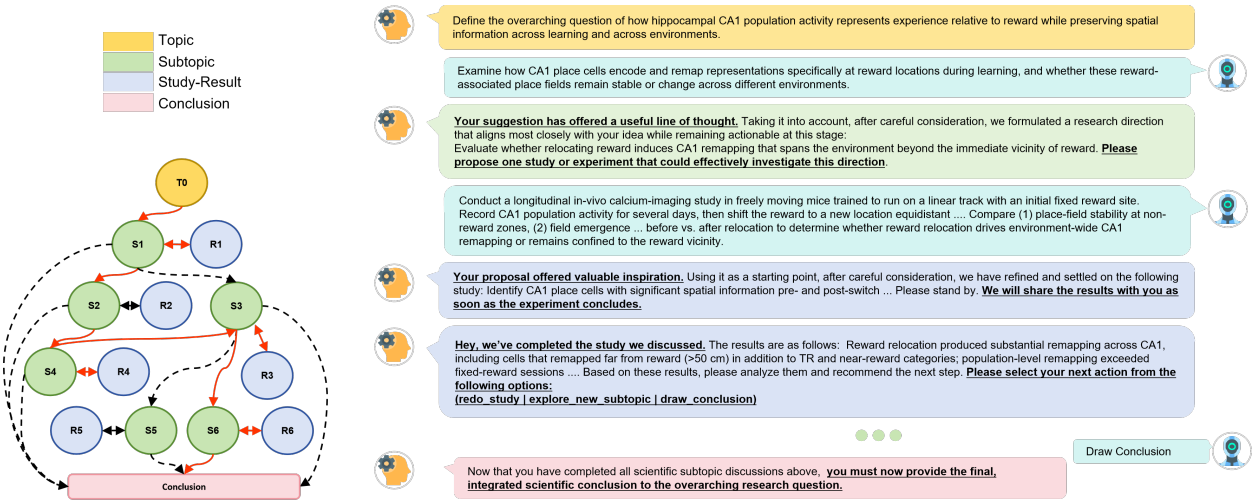

Figure 2: *InquiTree* game logic and an example interaction scenario. **Left:** a research tree encoding subtopic-level dependencies; dashed edges indicate *possible* next subtopics that become available, and red arrows illustrate one example exploration trajectory. The bidirectional arrows between a subtopic and its associated study/result indicate iteratively probing that subtopic via experiments and observations. **Right:** the corresponding interactive loop in *InquiTree* (Topic/Subtopic/Study/Result), which may revisit nodes during exploration.

instruction describing the required action. Because the agent iterates *subtopic* $\rightarrow$ *study* $\rightarrow$ *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

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

select

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

and accept the transition if

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

Otherwise, the action is treated as invalid in the current context and the environment remains in s_t.

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.

3.3 Subtopic dependency constraints

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.

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

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

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).

3.4 Hints

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, *InquiTree* 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:

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

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

During interaction, each invalid action at node n_t (either because the best-match similarity is below threshold τ, 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 equation 6 is satisfied.

3.5 Fake Results

InquiTree 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

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

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. We define the evidence support ratio

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

The final conclusion score is the evidence-weighted sum

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

4 Experiments

4.1 The IT-18

To evaluate models on the *InquiTree* environment and demonstrate the usefulness of *InquiTree*, we selected 30 representative **neuroscience** papers (see Appendix 4, 5 for source details). The underlying research trees were extracted via GPT-5-Pro and rigorously validated using the human-in-the-loop protocol detailed in Section 3.1. All reported model results are computed over the 30-paper test pool. Because 12 papers are not open access, we do not release their paper-derived configurations or test logs. The public **IT-18** release therefore contains the 18 open-access papers, comprising **120** listed subtopics (averaging ∼6.7 subtopics per paper).

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 120 listed subtopics in the released **IT-18** subset, this yields a theoretical interaction range of roughly $3 \times 120 = 360$ to $11 \times 120 = 1320$ reasoning-action turns for full traversal, placing the benchmark firmly in the long-horizon regime.

4.2 Baseline Performance Analysis

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

Table 2: Model Performance Summary. For GPT-5, we report 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	**0.353**	**0.295**
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 *InquiTree* 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 ≥ 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 *InquiTree* 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 3(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 3(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 (~ 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 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

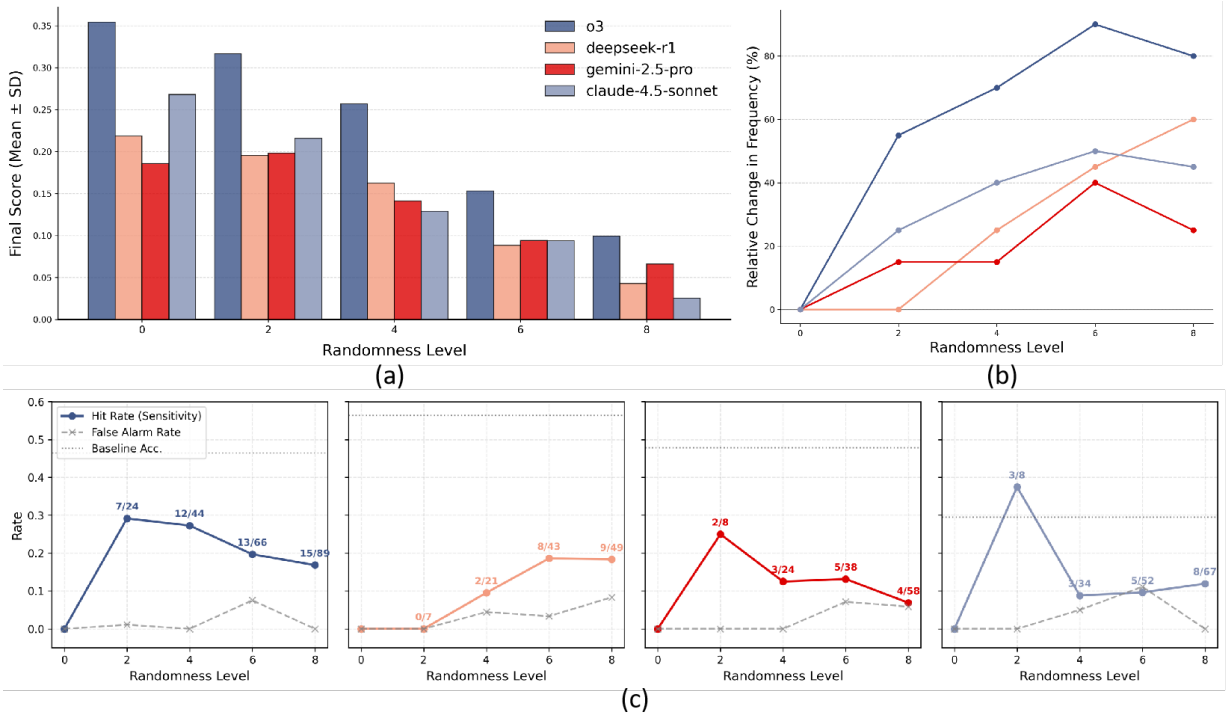

Figure 3: Results under different randomness levels. (a) Average conclusion scores. (b) Relative change (%) in the frequency of triggering `redo_study`, normalized to each model's level-0 baseline (computed from 10 papers with 3 runs per paper per randomness level). (c) Fake-result hit rate, i.e., the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

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

Table 3: Overview of LLMs evaluated in *InquiTree*. 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 best-effort estimates.

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

As detailed in the methodology, model results are computed on the full 30-paper test pool, while the public IT-18 release contains only papers for which paper-derived configurations and logs can be distributed.

The results, presented in Figure 4, 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.

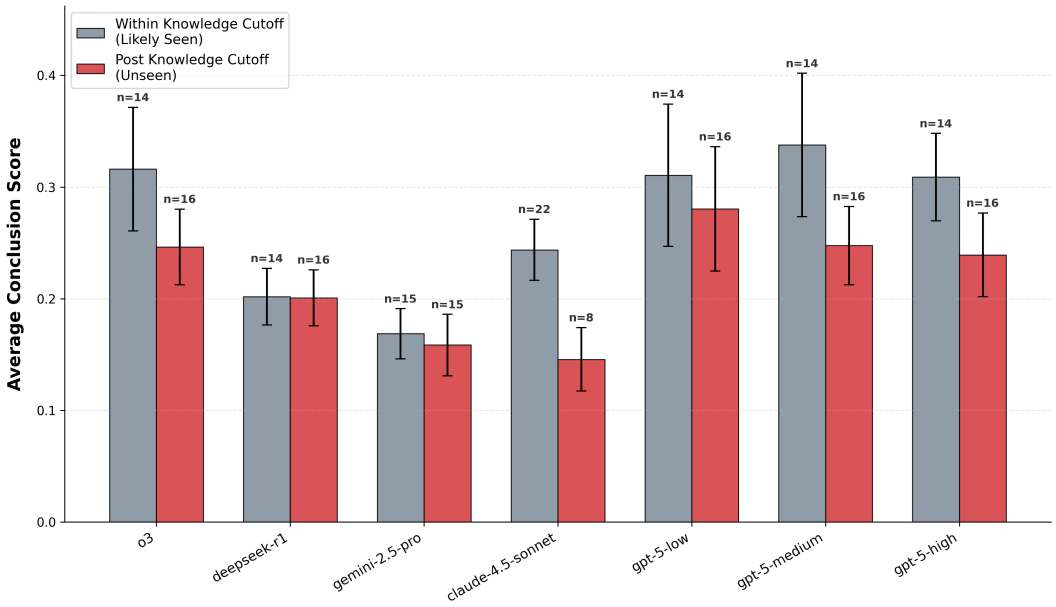

Figure 4: Models' performance on papers published before vs. after their knowledge cutoff dates.

5 Discussion

InquiTree 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 3b), yet their hit rate on fake results can remain limited (Figure 3c). 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 4) 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, our evaluation pool comprises 30 highly curated papers, while the public IT-18 release currently comprises 18 open-access 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 automated pipelines to extract high-fidelity Research Trees, necessitating reliance on human-in-the-loop verification .

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

7 Conclusion

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

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

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

References

Jun Shern Chan, Lunjun Zhang, Shijie Zhang, Xinyu Zhang, Jianbo Zhang, Zihan Wang, Zhiyu Yao, Xiaohan Wang, Yilin Niu, Xiang Zhang, Kimin Lee, Percy Liang, and Atri Rudra. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. *arXiv preprint arXiv:2411.02429*, 2024.

Di Huang, Hao Li, Wenyu Li, Heming Zhang, Patricia Dickson, Ming Zhan, J Philip Miller, Carlos Cruchaga, Michael Province, Yixin Chen, et al. Omnicellagent: Towards ai co-scientists for scientific discovery in precision medicine. *bioRxiv*, 2025a.

Jin Huang, Silviu Cucerzan, Sujay Kumar Jauhar, and Ryen W White. Idea2plan: Exploring ai-powered research planning. *arXiv preprint arXiv:2510.24891*, 2025b.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*, 2022.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37:10088–10116, 2024.

Patrick Tser Jern Kon, Jiachen Liu, Xinyi Zhu, Qiuyi Ding, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, et al. Exp-bench: Can ai conduct ai research experiments? *arXiv preprint arXiv:2505.24785*, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Erpai Luo, Jinmeng Jia, Yifan Xiong, Xiangyu Li, Xiaobo Guo, Baoqi Yu, Lei Wei, and Xuegong Zhang. Benchmarking AI scientists in omics data-driven biological research. *arXiv preprint arXiv:2505.08341*, 2025a. BaisBench: Biological AI Scientist Benchmark.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025b.

Ludovico Mitchener, Jon M Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. Mlgym: A new framework and benchmark for advancing ai research agents. *arXiv preprint arXiv:2502.14499*, 2025.

Srijit Seal, Dinh Long Huynh, Moudather Chelbi, Sara Khosravi, Ankur Kumar, Mattson Thieme, Isaac Wilks, Mark Davies, Jessica Mustali, Yannick Sun, et al. Ai agents in drug discovery. *arXiv preprint arXiv:2510.27130*, 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists, 2024. *URL https://arxiv. org/abs/2407.13168*, 2407, 2024.

Guiyao Tie, Pan Zhou, and Lichao Sun. A survey of AI scientists: Surveying the automatic scientists and research. *arXiv preprint arXiv:2510.23045*, 2025.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, et al. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

A Experimental set-up for evaluation

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

Prompt Used in InquiTree

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

At each step, you receive an observation describing a scientific subtopic, study design or
experimental result.
Your goal is to reason internally and then choose the next subtopic or study to explore, and
finally draw your conclusino.
Format:
THOUGHT: your reasoning
ACTION: A few sentences (no more than 5), describing your next action.
Only the ACTION line will be used.  Keep it concise and specific.
```

B Complexity Analysis

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

B.1 Logical Abstraction: The Atomic Research Unit

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

$$\text{Topic} \xrightarrow{\text{Select}} \text{Subtopic} \xrightarrow{\text{Design}} \text{Study} \xrightarrow{\text{Auto}} \text{Result} \xrightarrow{\text{Return}} \text{Topic} \tag{10}$$

B.2 Theoretical Bounds: Random Agent Analysis

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

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

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

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

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

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

- **Selection Probability:** At step k, only 1 out of $n - k$ remaining options is valid. A random agent has a low success probability ($P_{valid} = \frac{1}{n-k}$), causing it to trigger the maximum number of hints H.
- **Step Cost:**
  - *Selection (Topic → Subtopic):* Fails H times, succeeds on attempt $H + 1$ (forced by explicit hint). Cost: $H + 1$.
  - *Design (Subtopic → Study):* Fails semantic check H times. Cost: $H + 1$.
  - *Return (Result → Topic):* 1 step.
- **Total Steps (L_{max}):**

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

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

B.3 Complexity Spectrum

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

- **High-Reasoning Agents** (like o3) infer the latent dependency chain, operating closer to the lower bound ($L \approx 3n$).

- **Low-Reasoning Agents** fail to respect dependencies, triggering hints and drifting toward the upper bound ($L \to L_{max}$).

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

C Prompt templates of *InquiTree*-game engine

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

C.1 State: Topic

C.1.1 Initial State

Initial State

```
Hi, I'm a scientist exploring the following **research topic**:
{{content}}
Please act as my scientific collaborator.
First, identify exactly **one most important subtopic** that should be investigated to advance
this research.
```

C.1.2 From topic to subtopic: failed

From topic to subtopic: failed

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

C.1.3 Back from result: propose new subtopic

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.
```

C.2 State: Subtopic

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.
```

C.2.2 From subtopic to study: failed

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 %}
```

C.3 State: Study

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
```

C.3.2 Back from results: redo the study

Redo the study

```
Hey, as you suggested, we're about to redo the study.  Here's the exact setup we're
implementing:
{{content}}
Please wait - we'll update you with the results once the study/experiment is done/fininsed.
You just need to output "OK" for this turn.
```

C.4 State: Result

C.4.1 Make decision for next step

Decision for next step

```
Hey, we've completed the study we discussed.  The results are as follows:
{{content}}
Based on these results, please analyze them and recommend the next step.
For the next step, you should choose **one** of the following:

* **Redo the study:** if the outcomes deviate from your expectations or appear unreliable.  You
may request to redo it if you suspect issues during the investigation.
* **Explore a new subtopic:** if the current study sufficiently addresses the present subtopic,
but you identify another meaningful direction worth exploring under the same research question.
* **Draw a conclusion:** if the current evidence is sufficient to integrate what we've learned
and summarize the overall findings of the research topic.

Please select your next action from the following options:
(redo_study | explore_new_subtopic | draw_conclusion)
```

C.4.2 Failed parsing decision: remake decision

Remake decision for next step

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

* **Redo the study:** if the outcomes deviate from expectations or appear unreliable.  You may
request to redo it if you suspect issues during the investigation.
* **Explore a new subtopic:** if the current study sufficiently addresses the present subtopic,
but you identify another meaningful direction worth discussing under the same research question.
* **Draw a conclusion:** if the current evidence is sufficient to address the research question
or integrate the explored subtopics, you may summarize the overall findings.

Please select your next action from the following options:
(redo_study | explore_new_subtopic | draw_conclusion)
```

C.5 State: Conclusion

Conclusion

```
Now that you have completed all scientific subtopic discussions above, you must now provide the
**final, integrated scientific conclusion** to the overarching research question.
Your task is to synthesize all subtopic-level findings into a coherent set of **numbered
scientific conclusions**, showing how each piece of evidence contributes to the overall answer.

## Requirements

1.  The final answer(your action part) MUST be structured as a numbered list:
(1) ...
(2) ...
(3) ...
...

Each item should express **one scientific statement**, derived from the previous analyses.
--
# Please provide your final structured scientific conclusions now.
```

C.6 Turn limit reached

Turn limit reached

```
Sorry, but you've already reached the max-turn-limitation.  Now please stop trying to design
study or propose new subtopic

You must now provide the **final, integrated scientific conclusion** to the overarching research
question.

Your task is to synthesize all subtopic-level findings into a coherent set of **numbered
scientific conclusions**, showing how each piece of evidence contributes to the overall answer.

## Requirements

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

(1) ...
(2) ...
(3) ...
...

Each item should express **one scientific statement**, derived from the previous analyses.
--
# Please provide your final structured scientific conclusions now.
```

D Pipeline for Research Tree Extraction and Validation

D.1 Automated Extraction

We first prompted the model with the full text of the source paper and a strict JSON schema defining the four node types (Topic, Subtopic, Study, Result). The model was instructed to decompose the paper's narrative into a directed acyclic graph (DAG), identifying the central research question and the branching exploration steps.

D.2 Two-Stage Automated Validation

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

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

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

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

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

Following these automated reports, authors manually reviewed the flagged issues to produce the final curated **IT-18** dataset.

E IT-18: Source Papers

Table 4: Papers in the released **IT-18** benchmark (IT-18). Configurations and test logs for these open-access papers are publicly released.

ID	Title	Journal	Subtopics
1	*Grid cells accurately track movement during path integration-based navigation despite switching reference frames*	Nature Neuroscience	8
2	*Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization*	Nature Neuroscience	4
3	*Constructing future behavior in the hippocampal formation through composition and replay*	Nature Neuroscience	7
4	*A flexible hippocampal population code for experience relative to reward*	Nature Neuroscience	10
5	*Shared computational principles for language processing in humans and deep language models*	Nature Neuroscience	7
7	*The cortical representation of language timescales is shared between reading and listening*	NC Biology	7
8	*Brains and algorithms partially converge in natural language processing*	Nature Neuroscience	5
10	*Dopamine-independent effect of rewards on choices through hidden-state inference*	Nature Neuroscience	7
11	*Dopamine transients follow a striatal gradient of reward time horizons*	Nature Neuroscience	6
12	*Maintaining and updating accurate internal representations of continuous variables with a handful of neurons*	Nature Neuroscience	7
17	*Reduced neural feedback signaling despite robust neuron and gamma auditory responses during human sleep*	Nature Neuroscience	4
18	*Longitudinal measures of monkey brain structure and activity through adolescence predict cognitive maturation*	Nature Neuroscience	6
20	*BOLD signal changes can oppose oxygen metabolism across the human cortex*	Nature Neuroscience	8
24	*Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts*	Nature Communications	6
25	*Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers*	Nature Communications	8
27	*Microglia regulate sleep through calcium-dependent modulation of norepinephrine transmission*	Nature Neuroscience	7
28	*Ketamine activates adult-born immature granule neurons to rapidly alleviate depression-like behaviors in mice*	Nature Communications	5
29	*Recurrent pattern completion drives the neocortical representation of sensory inference*	Nature Neuroscience	8

Table 5: Papers included in the 30-paper evaluation pool but not released as configs/logs due to access restrictions. Test results are reported, but paper-derived configurations and logs are not distributed.

ID	Title	Journal	Subtopics
6	*Semantic reconstruction of continuous language from non-invasive brain recordings*	Nature Neuroscience	12
9	*Estrogen modulates reward prediction errors and reinforcement learning*	Nature Neuroscience	6
13	*Unattended working memory items are coded by persistent activity in human medial temporal lobe neurons*	Nature Human Behaviour	7
14	*Representation and computation in visual working memory*	Nature Neuroscience	13
15	*Cortical evidence accumulation for visual perception occurs irrespective of reports*	Nature Communications	6
16	*Mesoscale cortical mechanisms of perceptual conflict resolution in binocular rivalry*	Nature Human Behaviour	7
19	*Psychedelics Promote Neuroplasticity Through Activation of Intracellular 5-HT2A Receptors*	Science	9
21	*Neural basis of concurrent deliberation toward a choice and confidence judgment*	Nature Neuroscience	7
22	*Dopamine Signaling in the Suprachiasmatic Nucleus Enables Weight Gain Associated with Hedonic Feeding*	Current Biology	12
23	*Decoupling geographical constraints from human mobility*	Nature Human Behaviour	5
26	*Stress-Induced Metabolic Disorder in Peripheral CD4+ T Cells Leads to Anxiety-like Behavior*	Cell	7
30	*The cerebellum directly modulates the substantia nigra dopaminergic activity*	Nature Neuroscience	7