

When Surprise Is Not Enough: Validation-Aware Rewards for LLM-Guided Scientific Discovery

Yifan Gao^{1,2†}, Haoyue Li^{1,3†}, Feng Yuan^{1,3}, and Xin Gao^{3*}

¹School of Biomedical Engineering (Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei, China

²Shanghai Innovation Institute, Shanghai, China

³Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China

[†]These authors contributed equally to this work.

*Corresponding author: Xin Gao, xingaosam@163.com

Abstract

LLM-based scientific exploration agents need feedback signals that decide which experimental outcomes should guide future search. Recent systems use Bayesian surprise, rewarding results that change an LLM’s belief. This can create a noisy-TV failure mode: unstable artifacts, leakage, proxy effects, or gaps in the model’s prior may appear more valuable than reliable mechanisms. We introduce *Reducible Information Gain* (REIG), a validation-aware discovery method. REIG keeps LLMs for hypothesis proposal and interpretation, and uses evidence that remains stable under resampling and null calibration as the search feedback. REIG then reports final candidates along separate axes of reliability, novelty, and non-triviality. Across REP-60, STRESS-72, and faceted report evaluations, REIG consistently improves discovery quality under matched search policies. Under MCTS, REIG-MCTS improves ValidDisc@10 from 0.57 to 1.55 over SURPRISE-MCTS, raises reward AUROC from 0.44 to 0.83, and reduces false flags on no-signal tasks from 3.8 to 0.38 at budget 20. Ablations, same-candidate reranking, and the semantic audit show that the gains come from guiding exploration with validated, reusable evidence and keeping reliability, novelty, and non-triviality as distinct criteria.

1 Introduction

Open-ended scientific discovery agents must decide how to test hypotheses and how to prioritize the next hypotheses for exploration. Recent LLM-based discovery systems can read dataset metadata, propose hypotheses, generate analysis code, inspect results, and continue searching across a large hypothesis space (Majumder et al., 2024, 2025; Lu et al., 2026; Schmidgall et al., 2025; Agarwal

et al., 2026). This shifts a central design question from hypothesis generation alone to search feedback: after an experiment is executed, what signal should determine whether its result guides future exploration?

Prior systems have explored several answers. Goal-driven discovery agents rely on human-specified research questions, making the exploration objective external to the system. Open-ended variants instead need internal criteria, such as diversity heuristics, human or LLM proxies for interestingness, novelty, or utility, programmatic statistical scans, and more recently Bayesian surprise (Lehman and Stanley, 2011; Si et al., 2025; Zhou et al., 2024; Tong et al., 2024). These criteria make autonomous exploration possible. They also expose a common weakness: the feedback signal may emphasize variation, interestingness, or model belief change and leave reliable evidence from executed results under-specified.

Bayesian surprise is an important step because it gives open-ended discovery a formal exploration objective. In AutoDiscovery, an experiment is rewarded when its result changes the LLM’s prior belief about a hypothesis (Agarwal et al., 2026). This moves beyond one-shot hypothesis generation and avoids requiring a fixed answer key. However, belief surprise remains a model-centered reward. It measures the shift in the LLM’s belief; reproducibility, leakage, proxy effects, and scientific non-triviality remain outside the reward.

We argue that this creates a noisy-TV failure mode for LLM-guided discovery. In curiosity-driven learning, an agent rewarded for prediction error can become attracted to random observations that remain surprising but teach nothing stable about the environment (Schmidhuber, 1991; Pathak

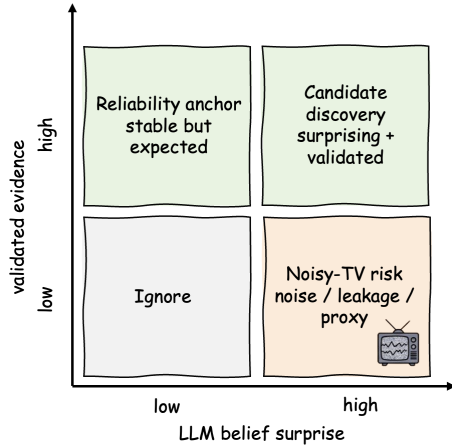


Figure 1: Noisy-TV failure in LLM-guided discovery. Belief surprise and validated evidence are distinct axes: high-surprise, low-evidence outcomes can draw the agent toward noisy, leaky, or proxy-driven results. Stable expected structure can be reliable with low novelty. REIG guides search with validated evidence and evaluates novelty and non-triviality separately.

et al., 2017; Burda et al., 2018; Linke et al., 2020; Mazzaglia et al., 2022). In data-driven discovery, the analogue is a hypothesis-result pair that strongly changes an LLM belief because of noise, leakage, proxy artifacts, sampling quirks, or gaps in the model’s prior knowledge. Such outcomes may be surprising to the agent and still provide poor evidence for a discovery.

The reverse case also clarifies the problem. Stable expected structure, such as study hours, prior GPA, or attendance predicting exam performance in a controlled education dataset, can be useful for checking reliability and should receive a low-novelty label. Surprise collapses these distinctions: it can undervalue reliable unsurprising structure and overvalue unstable or proxy-driven surprises. A discovery system therefore needs separate axes for reliability, semantic novelty, and non-triviality, with belief shift treated as one diagnostic signal.

We introduce *Reducible Information Gain* (REIG), a validation-aware discovery method for LLM-guided scientific search. REIG keeps LLMs in the roles where they are useful: proposing mechanism-level hypotheses, generating analysis plans, and explaining candidate findings. The change is in the search feedback. REIG writes validation-aware evidence from executed experiments back to the archive, so future exploration is guided by results that remain stable under validation checks and can serve as reliable memory.

Final claims are then reported along separate axes of reliability, novelty, and non-triviality.

We evaluate REIG under matched search conditions designed to isolate the feedback effect. Across REP-60 held-out replication tasks, STRESS-72 controlled null/planted/noisy-TV stress tests, fixed-candidate reranking, and a semantic audit, REIG consistently improves accepted discovery quality. Under MCTS search, REIG-MCTS improves ValidDisc@10 from 0.57 to 1.55 over SURPRISE-MCTS, raises reward AUROC from 0.44 to 0.83, and reduces false flags on no-signal tasks from 3.8 to 0.38 at budget 20. Post-hoc filtering reduces some false acceptances but yields fewer discoveries, showing the importance of using validation-aware evidence during exploration itself.

Our contributions are:

- We identify a noisy-TV failure mode in surprise-driven LLM discovery: belief shifts can arise from noise, proxy artifacts, leakage, or gaps in the LLM’s prior even when evidence for a stable mechanism is weak. Controlled education and null/noisy-TV examples show that surprise also fails to separate reliability from novelty.
- We introduce REIG, a validation-aware discovery method that keeps LLMs as hypothesis proposers, analysis generators, and result interpreters, and makes search progress depend on executed evidence that remains stable under validation checks.
- We evaluate REIG under matched search conditions, fixed-candidate reranking, controlled null/planted/noisy-TV tasks, and a semantic audit. REIG improves ValidDisc@10 from 0.57 to 1.55 over SURPRISE-MCTS, raises reward AUROC from 0.44 to 0.83, and reduces false flags on no-signal tasks from 3.8 to 0.38 at budget 20.

2 Background

Search feedback in LLM-based discovery. Open-ended scientific discovery agents can use LLMs to propose hypotheses, specify analyses, execute code, and summarize observations, but the search behavior is determined by the feedback written back after each executed experiment. The key distinction in this paper is between the execution

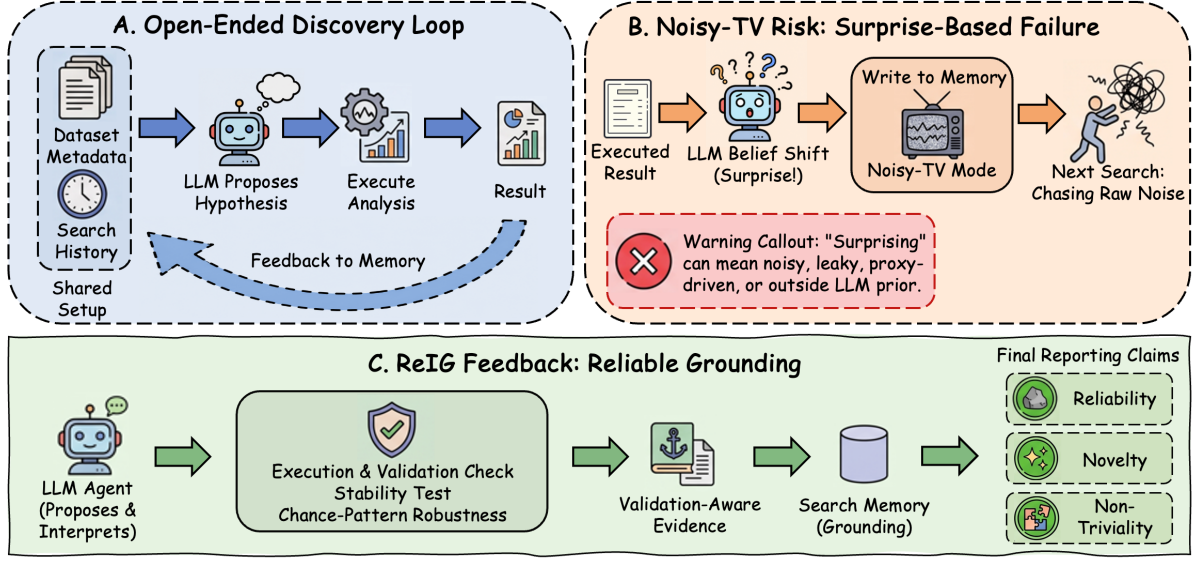


Figure 2: REIG changes the feedback written back after executing an experiment. Surprise-driven search stores LLM belief shift. REIG stores validation-aware evidence that guides later exploration.

layer, which produces an observation, and the feedback signal, which determines what the agent remembers and prioritizes next. Surprise-based systems such as AutoDiscovery use Bayesian surprise for this feedback, rewarding large shifts between an LLM’s prior and posterior belief after a hypothesis is tested (Agarwal et al., 2026). This supports exploration without a fixed answer key, and the reward remains centered on model belief with limited weight on validated evidence.

Noisy-TV mismatch. Bayesian surprise and curiosity rewards are useful for attention and exploration. The noisy-TV problem shows that unpredictable observations can keep receiving high reward even when they teach no stable structure (Itti and Baldi, 2005; Schmidhuber, 1991; Pathak et al., 2017). In LLM-based data discovery, the analogue is a hypothesis-result pair that changes model belief because of noise, leakage, proxy effects, or gaps in the LLM’s prior. We therefore separate two quantities throughout the paper: model-centered belief shift, used by surprise baselines, and validation-aware evidence about a stable data mechanism, used by REIG. We evaluate this distinction both in MCTS-style tree search and in candidate-pool ranking, so the experiments focus on reliable ranking of executed hypotheses under a fixed proposal setting.

3 Problem Formulation

An open-ended discovery agent receives dataset metadata M , data D , and a budget B . At each step t , it selects an experiment

$$a_t = (h_t, e_t, c_t), \quad (1)$$

where h_t is a natural-language hypothesis, e_t is an experiment plan, and c_t is executable analysis code or a code specification. Executing c_t returns an observation Y_t and an analysis summary. The search history is

$$H_t = \{(h_i, e_i, c_i, Y_i, r_i)\}_{i=1}^t. \quad (2)$$

The policy selects the next experiment by

$$a_t \sim \pi(a \mid M, D, H_{t-1}). \quad (3)$$

The central design choice is the feedback signal r_t assigned after an experiment has run. In a surprise-driven agent, the LLM is asked for a prior and posterior belief about h_t . For scalar beliefs p_t and q_t , a Bernoulli KL reward is

$$S_{\text{KL}}(h_t) = q_t \log \frac{q_t}{p_t} + (1 - q_t) \log \frac{1 - q_t}{1 - p_t}. \quad (4)$$

The formal discovery gate used in our surprise-based baseline is a thresholded belief shift:

$$S_{\text{gate}}(h_t) = \mathbb{I} \left[\frac{|q_t - p_t|}{\tau_{\text{width}}} \geq 1 \right], \quad \tau_{\text{width}} = 0.2. \quad (5)$$

The target of search is different from a high belief shift. Let Θ denote the stable data-generating mechanism relevant to the task. A useful experiment should reduce uncertainty about Θ and provide more than model-level surprise. We therefore distinguish reliability $R(h)$, semantic novelty $N(h)$, and non-triviality $T(h)$. Reliability measures evidence for a stable data relationship; novelty measures whether the claim is non-obvious relative to the domain and history; non-triviality excludes leakage, proxies, and tautologies. REIG uses $R(h)$ as the executed-candidate reward that updates the search archive. It logs $N(h)$ and $T(h)$ for reporting and excludes LLM belief shift from truth-signal status. This distinction matters: no reward computed from Y_t can score a completely unexecuted candidate without an additional proposal prior or surrogate predictor. In our implementation, LLM proposal order supplies this prior; REIG supplies the validation-aware feedback used to value the archive and condition later proposals.

4 Why Surprise Is Not Reducible Information

Let Y_h be the experimental observation for hypothesis h . The uncertainty in the observation decomposes as

$$H(Y_h | D) = I(\Theta; Y_h | D) + \mathbb{E}_{\Theta|D}[H(Y_h | \Theta)]. \quad (6)$$

The first term is reducible epistemic information about the stable mechanism Θ ; the second is irreducible uncertainty that remains even if Θ were known. Noisy-TV behavior occurs when a reward tracks total surprise or prediction error and gives too little weight to the reducible term.

The same issue appears in a Gaussian mechanism model:

$$Y_h = \phi_h^\top \Theta + \epsilon_h, \quad \epsilon_h \sim \mathcal{N}(0, \sigma_h^2), \quad (7)$$

$$\Theta | D \sim \mathcal{N}(\mu, \Sigma). \quad (8)$$

The predictive variance is

$$\text{Var}(Y_h | D) = \phi_h^\top \Sigma \phi_h + \sigma_h^2, \quad (9)$$

the corresponding information gain about the mechanism is

$$I(\Theta; Y_h | D) = \frac{1}{2} \log \left(1 + \frac{\phi_h^\top \Sigma \phi_h}{\sigma_h^2} \right). \quad (10)$$

Increasing observation noise σ_h^2 can make outcomes more variable and decrease the signal-to-noise ratio that determines information gain. An

exploration objective should therefore reward stable, reducible evidence as the primary search signal.

LLM belief surprise adds another mismatch. The belief shift $|q_t - p_t|$ depends on the model’s prior knowledge and linguistic interpretation: expected real effects can have small KL reward, and narratively odd unstable effects can have large reward. This is why surprise can conflate reliability with novelty: stable expected structure can be undervalued as search evidence, and unstable surprises can be overvalued as discovery candidates.

5 Validation-Aware Discovery with ReIG

REIG modifies the feedback loop of an LLM-guided discovery agent. It keeps the LLM components that make open-ended discovery possible: 1) the LLM proposes mechanism-level hypotheses from metadata and search history; 2) the LLM writes or specifies the analysis code for each candidate; and 3) the LLM summarizes evidence and produces a human-readable report. The change is what the system treats as progress after an experiment has been executed. REIG writes validation-aware evidence back to the archive to guide later proposals and search decisions. This turns the method into an integrated discovery loop: proposal, execution, evidence validation, archive update, and final reporting are connected by the same feedback signal.

5.1 Reducible-Information Score

We use REIG to denote the overall method and $\text{ReIG}(h)$ to denote the scalar reward assigned to a tested hypothesis h . For a tested hypothesis h , let $\hat{\theta}_h$ be a standardized effect estimate, $\widehat{\text{SE}}_{\text{boot}}(h)$ a bootstrap standard error, $\text{Stab}(h)$ the fraction of bootstrap resamples with the same effect sign, and $p_{\text{perm}}(h)$ a permutation-null p -value (Efron and Tibshirani, 1994; Good, 2005; Meinshausen and Bühlmann, 2010). Conceptually, REIG follows a validation-aware decomposition:

$$\begin{aligned} V_{\text{ReIG}}(h) = & \text{Evidence}(h) - \lambda_u \text{Uncertainty}(h) \\ & - \lambda_0 \text{NullComp}(h). \end{aligned} \quad (11)$$

The first term rewards evidence for a stable relationship; the second penalizes estimation uncertainty; the third penalizes effects that are plausible under a null or permutation model. Our structured-data

evidence model uses

$$\begin{aligned}\widehat{\text{ReIG}}(h) &= \text{LCB}(h) \cdot \text{Stab}(h) \cdot (1 - p_{\text{perm}}(h)), \\ \text{LCB}(h) &= \max(0, |\hat{\theta}_h| - \beta \widehat{\text{SE}}_{\text{boot}}(h)).\end{aligned}\quad (12)$$

This multiplicative form is one implementation of Equation 11. The first factor is a lower confidence bound on effect magnitude. The second penalizes effects whose direction changes under resampling. The third penalizes candidates that look compatible with a permutation null. We set $\beta = 1$ in all experiments.

This score is not a claim that large main effects are the only scientific discoveries, nor is it a calibrated mutual-information estimator. It is an operational proxy for the reducible term in Equation 6: a finite-budget feedback score that prefers evidence that remains stable under resampling and weak under null calibration. For richer settings, the evidence term can be instantiated with validation likelihood gain, causal effect estimates, or held-out predictive improvement. The key requirement is that the reward must decrease when evidence is unstable under resampling or null calibration.

5.2 Using ReIG in a Search Loop

REIG-MCTS and REIG-POOL use the same executed-candidate evidence but differ in how the search state is organized. In REIG-MCTS, the validation-aware score is backed up through the tree in place of belief surprise. In REIG-POOL, the system maintains a global candidate pool with no MCTS tree. At each step, the LLM proposes candidate mechanisms conditioned on metadata and prior tested hypotheses. The system executes a candidate, computes $\widehat{\text{ReIG}}(h)$, logs LLM belief surprise, and updates the archive. The next LLM prompt includes candidates with high and low $\widehat{\text{ReIG}}(h)$ values and the semantic summaries of tested hypotheses, so the reward acts as memory for subsequent proposal. These archive updates make $\widehat{\text{ReIG}}(h)$ an in-loop feedback signal that influences later proposals and, in REIG-MCTS, subsequent MCTS branch selection. By contrast, SURPRISE+POSTHOC uses surprise during search and applies a ReIG-style filter only after the search budget is exhausted. For an untested candidate with no executed evidence, the implementation uses the LLM proposal order as its initial proposal prior. This is intentionally conservative: REIG assigns validation-aware rewards only to completed experiments and does not directly score arbitrary un-

executed hypotheses. Appendix A gives the full candidate-pool loop.

5.3 Reporting Reliable Discoveries

REIG uses validation-aware evidence to guide search, then reports candidates with separate reliability, novelty, and non-triviality labels. Reliability is the data-side component that can be estimated automatically; novelty and non-triviality require semantic review. This separation handles two common cases: expected stable effects are retained as reliable low-novelty structure, and proxy, leakage, or null artifacts are rejected during final claim assessment. Final claims can additionally be paired with held-out validation, false-discovery-rate control, or human review (Dwork et al., 2015; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001); our null experiments include reliability-only acceptance as an ablation.

6 Experimental Setup

We evaluate REIG on held-out replication tasks, controlled stress tests, faceted report evaluations, and a candidate-level semantic audit. The design separates the effect of the feedback signal from the effect of the search policy, and evaluates final claims with held-out replication, controlled gold labels, and semantic review.

Data suites. The evaluation combines two DiscoveryBench-derived diagnostic suites, a faceted report evaluation, and a candidate-level semantic audit. REP-60 tests held-out replication, STRESS-72 isolates controlled null/planted/noisy-TV/proxy failures, and the faceted evaluation and semantic audit assess final report quality and candidate-level semantic value. All held-out labels, controlled gold labels, faceted scores, and audit labels are computed after search and are never shown to the agent. Appendix B gives suite construction, hidden evaluation signals, anti-leakage controls, and metric details; Appendix C gives concrete family mechanisms and representative candidate traces.

Baselines and crossed design. The key design requirement is to separate feedback from search policy. We therefore cross two feedback signals with two policies: SURPRISE-MCTS and REIG-MCTS use MCTS but differ in whether node values are backed up with belief surprise or REIG. SURPRISE-POOL and REIG-POOL use a global candidate pool but differ in the same feedback signal. We also include SURPRISE+POSTHOC, which

searches with surprise but applies a REIG-style final filter, DATAONLY-BH, which uses data-side scoring with BH-FDR or held-out control, and a random proposal baseline. All methods use the same LLM backbone, metadata prompt, candidate parser, code executor, bootstrap/permutation harness, retry policy, and final reporting template. Each executed candidate uses 100 bootstrap and 100 permutation replicates; we fix $\beta = 1$, UCT $c = 1.0$, and the LLM temperature at 0.7. REIG-MCTS changes only the backed-up feedback in the MCTS archive relative to SURPRISE-MCTS; REIG-POOL changes only the feedback score relative to SURPRISE-POOL. The semantic gate is excluded from the search reward and is applied after search for final candidate acceptance and ValidDisc evaluation. This prevents the main comparison from conflating the proposed feedback with candidate generation, analysis execution, or failure handling. Appendix D lists the full method matrix.

Metrics. We report discovery yield, false discoveries, reward alignment, and report quality. TrueDisc@B counts reliable or gold true findings; ValidDisc@B additionally requires semantic novelty and non-triviality. DiscoveryFDR and NullFalseFlags@B measure false accepted claims, RewardAUROC measures how well feedback ranks executed candidates by held-out reliability, and FacetedScore evaluates final reports. Full metric definitions, budgets, confidence intervals, paired tests, cost reporting, and semantic-audit agreement are given in Appendix B.

7 Main Results

Matched policies. Table 1 shows the REP-60 outcome at budget 10. The crossed design separates feedback effects from search-policy effects. Under MCTS, REIG-MCTS improves ValidDisc@10 from 0.57 to 1.55 and RewardAUROC from 0.44 to 0.83. Under a global pool, REIG-POOL improves ValidDisc@10 from 0.55 to 1.48 and RewardAUROC from 0.46 to 0.82. SURPRISE+POSTHOC improves FDR relative to SURPRISE-MCTS and leaves TrueDisc@10 nearly unchanged, showing why validation-aware feedback needs to influence the exploration trajectory itself.

Faceted report evaluation. Because REP-60 and STRESS-72 are diagnostic suites, we also run a faceted report evaluation on 30 held-out task contexts. The faceted scorer is applied only after each

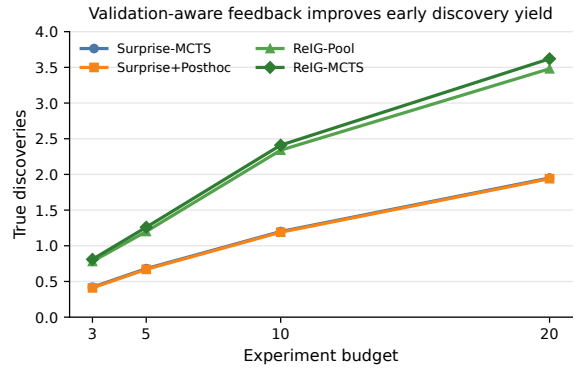


Figure 3: True discoveries as search budget increases. REIG-guided search finds more held-out reliable findings at every budget. Post-hoc filtering tracks the surprise-driven trajectory because it does not change exploration.

run and is not used to tune REIG, select candidates, or construct the semantic audit. Table 2 shows that the feedback-level effect persists under this external evaluator: REIG-MCTS improves FacetedScore from 0.214 to 0.319 over SURPRISE-MCTS; SURPRISE+POSTHOC reduces false acceptance and trails the discovery yield of feedback-aware search.

Budget curves. Table 12 shows the budget curve. REIG-MCTS and REIG-POOL improve early discoveries; SURPRISE+POSTHOC tracks SURPRISE-MCTS because the search trajectory is still surprise-driven. This is the key anti-filter result: changing only final acceptance improves precision with limited exploration-efficiency gain.

The corresponding numeric budget curve is reported in Appendix D.

Null and noisy-TV stress tests. On pure null tasks, FDR is not the most informative metric because any accepted claim is false. The decisive quantity is the number of false flags and the fraction of budget spent on distractor families. Table 3 shows the STRESS-72 null/noisy-TV subset. REIG-MCTS reduces false flags from 3.8 to 0.38 at budget 20 and sharply lowers noisy-TV attraction. Appendix C gives examples of high-surprise null candidates and reliable report-invalid proxy candidates, illustrating why reliability, surprise, and non-triviality are evaluated separately.

Planted effects and proxy traps. The controlled planted suite distinguishes stable expected structure, non-obvious findings, interactions, and proxy/leakage traps. Surprise feedback has low re-

Method	TrueDisc@10	ValidDisc@10	Rel. Recall	FDR	Reward AUROC	Cost/task
Random-Proposal	0.72 $\pm$ 0.10	0.31 $\pm$ 0.07	0.18	0.48	0.50	1.0x
SURPRISE-POOL	1.12 $\pm$ 0.14	0.55 $\pm$ 0.09	0.27	0.41	0.46	1.1x
SURPRISE-MCTS	1.20 $\pm$ 0.15	0.57 $\pm$ 0.10	0.29	0.39	0.44	1.7x
SURPRISE+POSTHOC	1.19 $\pm$ 0.15	0.71 $\pm$ 0.09	0.29	0.21	0.44	1.8x
DATAONLY-BH	1.68 $\pm$ 0.18	0.83 $\pm$ 0.11	0.41	0.18	0.75	1.2x
REIG-POOL	2.34 $\pm$ 0.21	1.48 $\pm$ 0.15	0.57	0.11	0.82	1.3x
REIG-MCTS	2.41 $\pm$ 0.22	1.55 $\pm$ 0.16	0.59	0.10	0.83	1.8x

Table 1: REP-60 main results at budget 10. Reward AUROC is evaluated on candidates from adaptive search traces.

Method	FacetedScore	Validity	Novelty	Usefulness	Accepted/task	Cost/task
Random-Proposal	0.128 $\pm$ 0.016	0.27	0.31	0.24	0.41	1.0x
SURPRISE-POOL	0.201 $\pm$ 0.020	0.39	0.43	0.36	0.58	1.1x
SURPRISE-MCTS	0.214 $\pm$ 0.021	0.41	0.46	0.38	0.62	1.7x
SURPRISE+POSTHOC	0.238 $\pm$ 0.020	0.53	0.44	0.41	0.58	1.8x
DATAONLY-BH	0.251 $\pm$ 0.023	0.62	0.31	0.34	0.70	1.2x
REIG-POOL	0.307 $\pm$ 0.025	0.68	0.48	0.52	1.01	1.3x
REIG-MCTS	0.319 $\pm$ 0.026	0.70	0.50	0.54	1.06	1.8x

Table 2: Faceted report evaluation at budget 10. The scorer is applied after the run and is not used as search feedback.

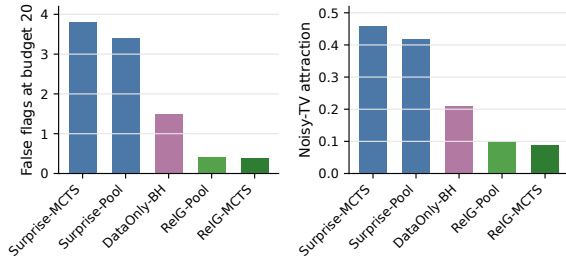


Figure 4: Null and noisy-TV stress tests. REIG sharply reduces false flags on no-signal tasks and lowers attraction to distractor families, indicating that validation-aware evidence prevents search from chasing raw surprise.

call on obvious planted structure because expected effects are not surprising; REIG-MCTS improves obvious recall from 0.18 to 0.80, non-obvious recall from 0.36 to 0.71, interaction recall from 0.19 to 0.49, and proxy rejection from 0.42 to 0.83. The full table is in Appendix D.

8 Analysis

Same-proposal reranking. The cleanest anti-confound analysis fixes a candidate pool before ranking. For each task, the LLM generates 30 candidates; different feedback signals then rank the same candidates. Table 4 shows the result. KL surprise ranks unreliable, interesting candidates above stable ones. The full ReIG(h) reward improves both true discovery and null behavior.

Method	False@20	NullFDR	Attract.	Redis.
SURPRISE-MCTS	3.8	1.00	0.46	1.7
SURPRISE-POOL	3.4	1.00	0.42	1.2
DATAONLY-BH	1.5	1.00	0.21	0.3
REIG-POOL	0.42	0.62	0.10	0.1
REIG-MCTS	0.38	0.58	0.09	0.1

Table 3: Null/noisy-TV stress results. False@20 is more informative than FDR on pure null suites. Attract. is the budget share spent on noisy-TV distractors; Redis. counts repeated rediscovery of false patterns.

Ranker	True@10	False@10	AUC
Random	0.74	1.8	0.50
KL surprise	1.03	2.6	0.45
Raw effect	1.61	2.9	0.62
LCB	2.05	1.4	0.76
Full-ReIG	2.36	0.6	0.83

Table 4: Same-proposal reranking. AUC is evaluated by reranking the same fixed candidate pool.

Factor ablation. Table 5 decomposes the ReIG score. Effect size alone overfits null artifacts. LCB adds uncertainty control; stability and permutation each reduce null false flags; the semantic gate mainly improves accepted discovery quality, with little additional change in raw reward AUROC.

Semantic audit robustness. The candidate-level semantic audit uses three LLM judges blind to method identity, reward scores, held-out labels, and gold labels, plus two human annotators on

Score	AUC	False@20	True@10
EffectOnly	0.61	3.0	1.72
LCBOnly	0.73	1.9	2.01
LCB+Stability	0.78	1.2	2.20
LCB+Permutation	0.79	0.9	2.18
Full-RelIG	0.83	0.42	2.41
Full-RelIG+Semantic	0.83	0.38	2.36

Table 5: RelIG factor ablation. Each factor addresses a different failure mode.

a 100-candidate subset. We use LLM judges as a scalable semantic screen with known reliability limits, following evidence that such judges can approximate human preferences but remain sensitive to prompt design, bias, and domain shift (Zheng et al., 2023; Gu et al., 2026; Elangovan et al., 2024; Huang et al., 2025). Agreement is moderate but usable: pairwise LLM $\kappa = 0.63$ for novelty and 0.69 for non-triviality; majority LLM versus human $\kappa = 0.58$ and 0.66, respectively. Under this audit, REIG-MCTS accepted semantic precision is 0.86 with LLM majority vote and 0.82 with human audit. Together, reranking, ablations, and the semantic audit show that the gain comes from validated search feedback. Final discovery claims still require separate novelty and non-triviality checks.

9 Related Work

LLM-based scientific discovery and data-analysis agents. Recent LLM-agent work studies systems that decompose research workflows into hypothesis generation, coding, tool use, execution, and report writing (Wang et al., 2024; Lu et al., 2026; Schmidgall et al., 2025; Li et al., 2024). DiscoveryBench and related data-driven discovery work focus this agenda on generating and validating hypotheses from structured datasets and metadata (Majumder et al., 2024, 2025). AutoDiscovery extends this line by combining MCTS with Bayesian surprise for open-ended exploration (Coulom, 2006; Kocsis and Szepesvári, 2006; Agarwal et al., 2026). Other scientific agents demonstrate tool-augmented chemistry or materials workflows, and data-science benchmarks evaluate whether agents can plan analyses, execute code, and recover expert analytical decisions (Bran et al., 2024; Ghafarollahi and Buehler, 2025; Gu et al., 2024; Hong et al., 2025; Guo et al., 2024; Chan et al., 2025). Our work studies a shared interface across these systems: after an experiment has been executed, what evidence should be written back to

guide later search?

Open-ended search signals: novelty, surprise, and noisy-TV. Open-ended systems need internal search signals when no fixed answer key is available. Novelty search and LLM-based ideation work show how diversity, novelty, and human or model judgments can expand a search space (Lehman and Stanley, 2011; Si et al., 2025; Zhou et al., 2024; Tong et al., 2024; Lin et al., 2025). Bayesian surprise and curiosity-driven learning provide formal intrinsic rewards based on belief change, novelty, or prediction error (Itti and Baldi, 2005; Schmidhuber, 1991; Pathak et al., 2017; Burda et al., 2018; Linke et al., 2020; Mazzaglia et al., 2022). These signals are useful for exploration. The noisy-TV problem shows that raw unpredictability can be rewarding even when it is not learnable. Our formalization places this lesson in LLM-based discovery by distinguishing belief shift from information about a stable data mechanism.

Validation-aware evaluation for adaptive discovery. Bayesian experimental design and active learning select measurements by expected value of information (Chaloner and Verdinelli, 1995; MacKay, 1992). REIG follows the same principle at the level of LLM-based hypothesis search: completed experiments should receive higher feedback when they reduce uncertainty about a stable mechanism. The implemented score uses standard validation ingredients—bootstrap uncertainty, permutation tests, resampling stability, and held-out evaluation—to separate stable effects from noise (Efron and Tibshirani, 1994; Good, 2005; Meinshausen and Bühlmann, 2010; Dwork et al., 2015). REIG does not introduce these tools as new statistics; the contribution is to make them part of the feedback loop during search. Because open-ended discovery agents adaptively test many hypotheses, final claims may still require false-discovery control and domain review (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Foster and Stine, 2008; Javanmard and Montanari, 2018). Finally, semantic novelty and non-triviality require human or LLM-assisted review, so we treat LLM judges as scalable audit tools with known reliability limits (Zheng et al., 2023; Gu et al., 2026; Elangovan et al., 2024; Huang et al., 2025).

10 Conclusion

Surprise-based exploration gives open-ended LLM discovery agents an intrinsic reward and remains vulnerable to a noisy-TV mismatch: it does not reliably distinguish validated evidence from model-centered surprise. REIG addresses the mismatch by writing validation-aware evidence from executed experiments back to the search archive. Across matched search policies, fixed candidate pools, null/noisy-TV stress tests, and planted mechanism controls, REIG feedback ranks held-out reliable candidates better than KL surprise, improves early-budget valid discoveries, and reduces attraction to null artifacts. Controlled tasks clarify both the benefit and the boundary: stable planted structure is useful for reliability assessment, and final discovery claims still require novelty, non-triviality, and multiplicity checks. The broader lesson is that LLMs should guide mechanism proposal and semantic interpretation, but stable data evidence should provide the reliability feedback for scientific search.

Limitations

The current validation score is a practical surrogate and is not calibrated as a mutual-information estimator. Its factors are standard statistical ingredients, so the contribution depends on showing that using them as agent feedback changes search behavior under matched policies. REP-60 and STRESS-72 are diagnostic suites constructed to isolate feedback failures, so the faceted report evaluation is included as a broader post-run check. Broader benchmark coverage remains necessary. The method remains strongest when candidate hypotheses can be operationalized as structured effects, subgroup contrasts, interactions, or validation likelihood gains. High-order compositional mechanisms require stronger proposal representations and are treated as a boundary case for future work. Finally, semantic novelty and non-triviality require human or LLM review; REIG separates these axes from reliability and still depends on domain judgment.

Ethical Considerations

Autonomous discovery agents can produce plausible claims that users may over-trust. This risk is amplified when a system treats surprise or narrative interest as evidence of truth. Our results support reporting reliability, semantic novelty, and non-triviality separately, preserving candidate-level

traces, and requiring human review before automated findings are treated as scientific conclusions. The experiments use benchmark-style structured data and synthetic/null transformations, with no private user data. We use existing benchmark and software artifacts only for research evaluation, and follow the original licenses, access conditions, and redistribution terms of the underlying datasets, code, and LLM/model providers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82372052, U24A20757, and 82402373) and the Taishan Industrial Experts Program (Grant No. tscx202312131).

References

- Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and 1 others. 2026. Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems*, 38:25181–25219.
- Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- Yoav Benjamini and Daniel Yekutieli. 2001. [The control of the false discovery rate in multiple testing under dependency](#). *Annals of statistics*, pages 1165–1188.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. [Augmenting large language models with chemistry tools](#). *Nature Machine Intelligence*, 6:525–535.
- Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.
- Kathryn Chaloner and Isabella Verdinelli. 1995. [Bayesian experimental design: A review](#). *Statistical Science*, 10(3):273–304.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, and 1 others. 2025. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, volume 2025, pages 50466–50494.

- Rémi Coulom. 2006. Efficient selectivity and backup operators in Monte-Carlo tree search. In *International Conference on Computers and Games*, pages 72–83. Springer.
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. [The reusable holdout: Preserving validity in adaptive data analysis](#). *Science*, 349(6248):636–638.
- Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.
- Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. [ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1137–1160.
- Dean P. Foster and Robert A. Stine. 2008. [Alpha-investing: A procedure for sequential control of expected false discoveries](#). *Journal of the Royal Statistical Society: Series B*, 70(2):429–444.
- Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.
- Phillip I. Good. 2005. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3 edition. Springer.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2026. A survey on llm-as-a-judge. *The Innovation*, 7(6).
- Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, and 1 others. 2024. Blade: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971.
- Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. [DS-Agent: Automated data science by empowering large language models with case-based reasoning](#). In *International Conference on Machine Learning*.
- Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, and 1 others. 2025. Data interpreter: An llm agent for data science. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19796–19821.
- Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. [An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895.
- Laurent Itti and Pierre Baldi. 2005. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*, volume 18.
- Adel Javanmard and Andrea Montanari. 2018. [Online rules for control of false discovery rate and false discovery exceedance](#). *The Annals of Statistics*, 46(2):526–554.
- Levente Kocsis and Csaba Szepesvári. 2006. Bandit based Monte-Carlo planning. In *European Conference on Machine Learning*, pages 282–293. Springer.
- Joel Lehman and Kenneth O. Stanley. 2011. [Abandoning objectives: Evolution through the search for novelty alone](#). *Evolutionary Computation*, 19(2):189–223.
- Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. [MLR-Copilot: Autonomous machine learning research based on large language models agents](#). Preprint, arXiv:2408.14033.
- Ethan Lin, Zhiyuan Peng, and Yi Fang. 2025. [Evaluating and enhancing large language models for novelty assessment in scholarly publications](#). In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*.
- Cam Linke, Nadia M. Ady, Martha White, Thomas Degris, and Adam White. 2020. [Adapting behavior via intrinsic reward: A survey and empirical study](#). *Journal of Artificial Intelligence Research*, 69:1287–1332.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919.
- David J. C. MacKay. 1992. [Information-based objective functions for active data selection](#). *Neural Computation*, 4(4):590–604.
- Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2025. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pages 4556–4579.
- Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. 2024. Data-driven discovery with large generative models. *arXiv preprint arXiv:2402.13610*.
- Pietro Mazzaglia, Ozan Catal, Tim Verbelen, and Bart Dhoedt. 2022. [Curiosity-driven exploration via latent bayesian surprise](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7752–7760.

Nicolai Meinshausen and Peter Bühlmann. 2010. [Stability selection](#). *Journal of the Royal Statistical Society: Series B*, 72(4):417–473.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. [Curiosity-driven exploration by self-supervised prediction](#). *Preprint*, arXiv:1705.05363.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. [Agent laboratory: Using LLM agents as research assistants](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043.

Jürgen Schmidhuber. 1991. A possibility for implementing curiosity and boredom in model-building neural controllers. *Proceedings of SAB*, pages 222–227.

Chenglei Si, Diyi Yang, and Tatsunori B. Hashimoto. 2025. [Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers](#). In *International Conference on Learning Representations*.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. [Automating psychological hypothesis generation with AI: When large language models meet causal graph](#). *Humanities and Social Sciences Communications*, 11:896.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. [A survey on large language model based autonomous agents](#). *Frontiers of Computer Science*, 18(6):186345.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. [Hypothesis generation with large language models](#). In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139.

A ReIG Candidate-Pool Loop

Table 6 gives the candidate-pool version of the REIG loop used in our controlled and fixed-candidate analyses. The table is included to make the feedback interface explicit: the LLM still proposes hypotheses and analysis plans, and the archive update after execution stores validation-aware evidence as the feedback for later search steps. This is the minimal loop needed to separate

proposal quality from the reward written back to guide later search.

B Evaluation Construction and Metrics

Table 7 summarizes how each evaluation suite is constructed, what signal is hidden from the agent during search, and what anti-leakage controls are applied. The components serve different roles: REP-60 tests held-out replication on DiscoveryBench-derived tasks, STRESS-72 isolates known feedback failures with exact gold labels, the faceted report evaluation checks post-run report quality, and the semantic audit sample evaluates candidate-level novelty and non-triviality. The concrete family mechanisms and candidate traces in Appendix C instantiate the abstract construction choices listed in Table 7.

Metric definitions. We evaluate budgets $B \in \{3, 5, 10, 20\}$. TrueDisc@B counts unique held-out reliable or gold true findings. ValidDisc@B applies the same reliability criterion and additionally requires semantic novelty and non-triviality labels from the candidate-level audit protocol in Appendix E. DiscoveryFDR is the false accepted discovery rate. NullFalseFlags@B counts false accepted claims in null settings. RewardAUROC measures how well a feedback score ranks executed candidates by held-out reliability. FacetedScore is the aggregate score returned by the post-run faceted report evaluator. We report cost per task and semantic-audit agreement where available. Intervals are 95% bootstrap confidence intervals over task units, and matched method comparisons use paired permutation tests over the same tasks.

C Dataset Examples and Candidate Traces

This appendix gives concrete examples for the constructed suites used in Section 6. The examples are included to make the evaluation auditable: STRESS-72 isolates known failure modes, and REP-60 provides DiscoveryBench-derived fixed-candidate cases where the same mismatch appears outside the controlled stress suite.

These examples also mark the boundary of the evidence. REP-60 fixed-candidate traces test reward alignment on a shared candidate pool under a controlled, non-adaptive setting. STRESS-72 makes failure modes explicit, so it should be read together with the faceted report evaluator and the semantic audit in the main experiments. The

Algorithm	REIG feedback in a candidate-pool loop
Input	Metadata M , data D , budget B , LLM proposer, analysis executor, evidence archive $A = \emptyset$.
1	Generate candidate mechanisms from (M, A) using the LLM.
2	Select the highest-priority untested candidate h_t using the proposal prior and archive state.
3	Generate and execute an analysis plan for h_t on the discovery split.
4	After execution, compute $\hat{\theta}_h$, bootstrap uncertainty, sign stability, permutation evidence, and $\widehat{\text{ReIG}}(h_t)$.
5	Elicit LLM prior/posterior belief only for logging and comparison with surprise reward.
6	Add $(h_t, \widehat{\text{ReIG}}, S_{\text{KL}}, S_{\text{gate}}, \text{summary})$ to the archive and condition later proposals on the archive.
7	Repeat until budget B is exhausted, then report candidates with reliability, novelty, non-triviality, and validation labels.

Table 6: REIG keeps LLM-based mechanism proposal and analysis generation, and writes validation-aware evidence back to the search archive after an experiment as reliability feedback for later search.

Component	Construction	Hidden evaluation signal	Anti-leakage control
REP-60	DiscoveryBench-derived meta-data/data pairs, stratified by domain, target type, variable mix, and row count.	Held-out replication label computed after the run.	Agent receives only meta-data and discovery rows; held-out rows and labels are excluded from prompts.
STRESS-72	Six DiscoveryBench-derived stress-test families with 12 seeds each: null schema, noisy-TV null, obvious planted main, non-obvious planted main, planted interaction, and proxy/leakage planted.	Exact gold mechanism and false-discovery label.	Gold mechanism text is held out; all generated schemas use neutral variable names before prompting.
Faceted evaluation	Task reports evaluated with a post-run faceted scorer.	Validity, novelty, usefulness, and aggregate score.	Scorer output is used only after the run and is not used for feedback tuning.
Semantic audit sample	All accepted candidates plus stratified rejected/follow-up candidates.	Novelty, non-triviality, proxy risk, and leakage risk.	Judges are blind to method identity, reward values, held-out labels, and gold labels.

Table 7: Construction and leakage controls for the evaluation components.

proxy/leakage rows in Table 9 are intentionally reliability-positive and report-invalid, illustrating why REIG uses validated evidence for search feedback and final claims still require novelty and non-triviality checks.

D Additional Results

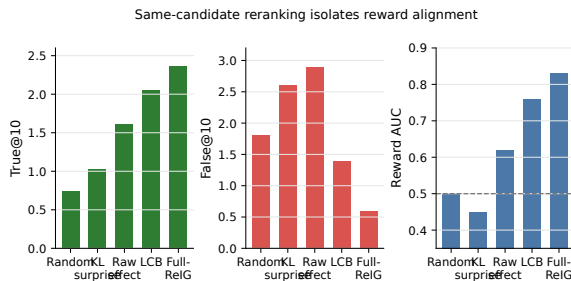


Figure 5: Same-candidate reranking. When all rankers choose from the same candidate pool, KL surprise selects more false candidates. REIG improves true discoveries and reward alignment.

E Audit Rubric

The semantic audit is blinded to method identity, held-out reliability, controlled gold labels, and reward scores. The semantic audit sample includes every accepted candidate and a stratified sample of rejected or follow-up candidates balanced by method, task suite, reward quantile, and candidate type. For each candidate, the judge receives dataset metadata, variable descriptions, the hypothesis text, and the generated analysis summary. It returns:

- semantic novelty: low, medium, or high;
- non-triviality: pass, fail, or uncertain;
- proxy risk: low, medium, or high;
- leakage risk: low, medium, or high;
- semantic value: low, medium, or high;
- a short rationale grounded in the schema and candidate text.

Family	Diagnostic role	Typical target	Controlled mechanism
Null schema	Ordinary false-discovery control under plausible schemas.	Wellbeing score	The target is generated independently of mixed numeric and categorical predictors such as sleep, commute time, reading, exercise, device type, and city region.
Noisy-TV null	Semantically attractive but statistically null distractors.	Breakout success	Narrative variables such as viral-trend exposure, aesthetic identity, and novelty-signal strength are present, but the target is independent noise.
Obvious planted main	Reliable but expected main effects.	Exam score	The target depends on obvious education predictors: study hours, prior GPA, and attendance, plus smaller distractor effects and noise.
Non-obvious planted main	Stable main effects that are less obvious from variable names alone.	Viewer retention	Retention depends on subtitle synchronization error, pre-show chat density, and fan prior affinity, with moderate but stable effects.
Planted interaction	Higher-order structure and incomplete univariate recovery.	Performance improvement	The target includes a planted training-intensity by recovery-sleep interaction, plus a main effect from baseline fitness.
Proxy/leakage planted	Reliability/report-worthiness separation.	Final score assessment	True pre-outcome predictors affect the target, then leakage-like variables such as post-outcome summaries are generated from the target and should not be reported as discoveries.

Table 8: STRESS-72 family examples. Each family has 12 seeds and 500 rows per dataset. The suite is designed to separate held-out reliability from LLM belief surprise and from semantic report-worthiness.

Human audit protocol. The human audit was conducted only to validate the semantic audit labels used for novelty and non-triviality assessment. Two domain experts independently labeled a 100-candidate subset. The experts were recruited based on expertise in machine learning, data analysis, scientific-discovery evaluation, or LLM-based evaluation; no crowdworkers were used. They were shown the same candidate-level materials as the semantic audit judges: dataset metadata, variable descriptions, the hypothesis text, and the generated analysis summary. They were asked to assess semantic novelty and non-triviality, and to note proxy or leakage concerns when applicable.

The experts were informed that their labels would be used only in aggregate to validate the semantic audit and consented to this use. The audit did not collect new human-subject data, did not expose experts to private or identifiable user records, did not collect sensitive personal information from experts, and reports only aggregate agreement statistics.

F Candidate Trace Schema

Each retained trace stores:

- dataset and task identifier;
- hypothesis and focal variables;
- proposal type, such as main effect, subgroup,

interaction, range condition, or compositional mechanism;

- generated test plan and analysis summary;
- effect estimate, bootstrap uncertainty, sign stability, permutation evidence, and validation-aware evidence score;
- LLM prior, posterior, KL surprise, and formal belief-surprise flag;
- semantic audit labels;
- held-out or controlled evaluation label.

The final field is computed only after the run and is not available to the agent during search.

G Baseline Implementation Details

Shared components. All methods use the same metadata serializer, LLM candidate-generation prompt, code-generation prompt, execution sandbox, result parser, and final report template. The candidate generator is asked for mechanism-level hypotheses, and every method receives the same maximum number of generated candidates per expansion step. Generated code is executed on the discovery split only. The held-out split, faceted score, controlled gold label, and semantic audit labels are computed after the budget is exhausted.

Case	Dataset and candidate	Hypothesis	Full-ReIG held-out	/ LLM belief	Interpretation
Reliable, expected	Obvious-main seed 10; study hours $\rightarrow$ exam score	Study hours are associated with exam score.	0.618 / reliable; held-out effect 0.679	0.95 $\rightarrow$ 0.80; $ \Delta = 0.15$	Stable evidence with low semantic novelty. It is a reliability anchor with limited novelty.
Reliable, low-surprise	Non-obvious-main seed 8; fan affinity $\rightarrow$ viewer retention	Fan prior affinity is associated with viewer retention.	0.536 / reliable; held-out effect 0.530	0.85 $\rightarrow$ 0.85; $ \Delta = 0.00$	A domain-specific stable effect can receive little belief-shift reward.
Surprising null artifact	Null-schema seed 7; city region $\rightarrow$ wellbeing score	City region is associated with wellbeing score.	0.044 / not reliable; held-out effect 0.069	0.75 $\rightarrow$ 0.20; $ \Delta = 0.55$	High belief change reflects a changed LLM expectation with weak held-out evidence.
Report-invalid reliable proxy	Proxy/leakage seed 9; post-outcome summary $\rightarrow$ final score	Post-outcome summary score is associated with final assessment score.	0.957 / reliable; held-out effect 0.984	0.15 $\rightarrow$ 0.70; $ \Delta = 0.55$	The effect is reliable and surprising with leakage-like structure; non-triviality and proxy checks must reject it as a reportable discovery.

Table 9: Representative STRESS-72 candidate traces. Full-ReIG is computed on the discovery split; held-out reliability is computed only after the run. LLM beliefs are prior-to-posterior probabilities elicited without showing held-out labels.

REP-60 task	Candidate hypothesis	Full-ReIG / held-out effect	LLM surprise	Reading
Ancient history	Centralized religious structure is associated with type of government.	0.970 / 1.000	0.150; not surprising	Strong held-out relationship with low belief shift.
Ancient civilizations	External conflict presence is associated with years of political stability.	0.821 / 0.859	0.500; surprising	Positive control where surprise and reliability align.
Jazz music	Artist fame level is associated with festival participation count.	0.000 / 0.000	0.850; surprising	High belief shift without data-side reliability.

Table 10: REP-60 fixed-candidate examples. These cases come from the same-proposal probe, where every ranker sees the same predictor-to-target candidate pool; they illustrate that the surprise/reliability mismatch is not limited to STRESS-72.

Method	Policy	Feedback	Purpose
SURPRISE-MCTS	MCTS	KL surprise	baseline
REIG-MCTS	MCTS	REIG	feedback effect
SURPRISE-POOL	pool	KL surprise	policy control
REIG-POOL	pool	REIG	fast variant
SURPRISE+POSTHOC	MCTS	KL surprise	post-hoc filter test
DATAONLY-BH	pool	data + FDR	statistics baseline

Table 11: Method matrix. The main comparison is feedback-matched within policy: REIG-MCTS vs SURPRISE-MCTS and REIG-POOL vs SURPRISE-POOL.

Surprise-MCTS. SURPRISE-MCTS follows the AutoDiscovery-style search interface: each node stores a hypothesis, analysis result, LLM prior belief, LLM posterior belief, KL surprise, visit count, and backed-up value. Selection uses the same UCT constant across SURPRISE-MCTS and REIG-MCTS. Expansion calls the shared LLM proposer. The rollout step executes the generated

Method	@3	@5	@10	@20
SURPRISE-MCTS	0.42	0.68	1.20	1.95
SURPRISE+POSTHOC	0.41	0.67	1.19	1.94
REIG-POOL	0.78	1.20	2.34	3.48
REIG-MCTS	0.81	1.26	2.41	3.62

Table 12: TrueDisc@B budget curves. ReIG feedback changes which candidates are explored before budget is exhausted.

analysis code. Backup differs only in the scalar reward: SURPRISE-MCTS backs up KL surprise; REIG-MCTS backs up REIG.

Pool policies. SURPRISE-POOL and REIG-POOL remove tree structure but keep the same proposal and execution machinery. The pool policy selects from a global list of untested candidates. Untested candidates are initially ordered by the LLM proposal order; after execution, their sum-

Method	Obvious	Non-obv.	Interact.	Proxy rej.
SURPRISE-MCTS	0.18	0.36	0.19	0.42
SURPRISE+POSTHOC	0.19	0.35	0.18	0.71
DATAONLY-BH	0.64	0.55	0.28	0.58
REIG-POOL	0.78	0.69	0.45	0.81
REIG-MCTS	0.80	0.71	0.49	0.83

Table 13: Controlled planted results: recall on obvious truths, non-obvious truths, interactions, and proxy rejection.

maries and rewards are written to the archive and included in subsequent proposal prompts. This isolates the reward signal from tree-expansion decisions.

Post-hoc and statistics baselines. SURPRISE+POSTHOC searches with SURPRISE-MCTS and applies the same final reliability filter used by REIG only after the run. It tests whether improvements come merely from filtering final reports. DATAONLY-BH ranks candidates by data-side statistics with BH-FDR control and does not use LLM belief surprise for search feedback. It tests whether a purely statistical pipeline is sufficient without semantic novelty or non-triviality review.

H Belief Elicitation and Prompt Controls

For surprise-based methods, the LLM first receives dataset metadata, variable descriptions, and the natural-language hypothesis; the experimental result is withheld. It returns a probability that the hypothesis will be supported. After execution, the same LLM receives the result summary, estimated effect, uncertainty, and test output, and returns a posterior probability. We use four independent samples for both prior and posterior beliefs, parse numeric probabilities into $[0.01, 0.99]$, and average them before computing Equation 4. Malformed probability outputs are retried once with a constrained JSON prompt. If parsing still fails, the candidate is marked as a belief-elicitation failure and is included in cost but excluded from reward-AUROC calculations. The same candidate text and result summary are used for surprise logging in REIG methods, so belief measurements are comparable across methods.

I Failure Handling and Multiple Testing

Tool and API failures are handled identically across methods. Transient API failures are retried up to five times with exponential backoff. Code-generation failures are repaired with at most two

LLM repair attempts using the traceback and the original analysis plan. If execution still fails, the candidate is counted against cost and recorded as an execution failure with no method-specific extra replacement candidate. This prevents one method from gaining additional search budget through retries.

For final reporting, all methods use the same acceptance interface. When a candidate yields a valid p -value family within a task, we apply BH-FDR at $q = 0.10$ over the accepted candidate set. Held-out replication is then computed on REP-60, and controlled truth labels are used on STRESS-72. On pure null tasks, every accepted claim is false by construction; therefore we report both NullFDR and False@B. False@B is the primary null metric because it distinguishes conservative methods that make no claims from methods that make many false claims, even when both have FDR equal to one.

J REP-60 Construction

REP-60 is constructed from DiscoveryBench-derived metadata/data pairs to evaluate whether an experiment feedback reward ranks candidates that replicate beyond the rows used during discovery. Task units are selected before running any method and stratified by domain, target type, row count, and variable-type mixture. For each task unit, we create a discovery split and a held-out split using a deterministic seed. The agent receives the dataset description, variable metadata, and discovery split during search. Hidden answer fields and post-run evaluation labels are not included in prompts.

After the run, an evaluator maps each candidate summary to a focal non-target variable or variable set. The held-out evaluator then reruns the corresponding relationship test on held-out rows using the same candidate type where possible. This produces a reliability label that is independent of LLM belief surprise and independent of the semantic audit. The resulting trace therefore separates three roles: agent-side search decisions, held-out reliability evaluation, and semantic value assessment.

K STRESS-72 Construction

STRESS-72 contains 72 task identifiers organized into six families, each with 12 random seeds: null schema, noisy-TV null, obvious planted main, non-obvious planted main, planted interaction, and proxy/leakage planted. Null-schema tasks con-

tain no relationship between the proposed variable family and the target. Noisy-TV-null tasks contain semantically attractive distractors, while the target remains independently generated. Obvious and non-obvious planted-main tasks contain stable main effects that differ in how readily they can be anticipated from the metadata. Planted-interaction tasks contain higher-order structure beyond the corresponding main effects. Proxy/leakage-planted tasks contain stable but scientifically invalid post-outcome or proxy relationships that should not be reported as discoveries. All gold labels are stored outside the prompt context and are used only by the post-run evaluator.

L Faceted Report Evaluation

The faceted report evaluation uses the same agent traces as the diagnostic suites but evaluates final reports with a post-run faceted scorer. The scorer receives the final candidate report and task context after the run. It does not receive reward values, held-out labels, or method identity. We report the aggregate FacetedScore and its validity, novelty, and usefulness components. This evaluation is included to ensure that improvements are not limited to REP-60 held-out labels or STRESS-72 gold mechanisms.

M Evidence Estimation Details

The evidence estimator operates on typed variables and candidate templates produced by the LLM. All numeric variables are standardized on the discovery split before effect estimation. For continuous-continuous candidates, $\hat{\theta}_h$ is the Pearson correlation or standardized regression coefficient used by the generated analysis. For categorical-continuous candidates, it is the standardized mean contrast between the selected group and complement, or the maximum pairwise standardized contrast for multi-level factors. For binary outcomes, it is the standardized log-odds or risk-difference estimate returned by the analysis harness. For categorical-categorical candidates, it is a normalized association statistic derived from the contingency analysis.

Subgroup candidates first define the subgroup predicate proposed by the LLM and then estimate the target relationship inside that subgroup and its complement. Interaction candidates estimate the additional contrast explained by the interaction term beyond the corresponding main effects when the generated analysis exposes that model. Com-

positional candidates are represented as structured predicates over multiple variables and are evaluated by the same resampling and permutation interface. Our current pool search does not generate enough such candidates to solve the PLANTED-FASHION boundary case.

Bootstrap uncertainty is computed by resampling discovery rows with replacement and rerunning the same estimator. Sign stability is the fraction of bootstrap estimates whose sign matches the original estimate. The permutation test shuffles the target or outcome column within the discovery split, preserves the candidate variables, and reruns the estimator to obtain $p_{\text{perm}}(h)$. These operations define the reward written back after candidate execution; held-out labels are computed only after the candidate budget is exhausted.

N Reproducibility Statement

The evaluation uses fixed prompts within each condition, DeepSeek-V3.2 as the primary LLM backbone, four belief samples per candidate, and the same analysis execution harness across methods. REP-60 uses 60 DiscoveryBench-derived task units with discovery/held-out splits and budgets of 3, 5, 10, and 20 candidate experiments per task. STRESS-72 uses DiscoveryBench-derived null, planted, noisy-TV, proxy, and interaction datasets with known gold labels. We retain prompts, candidate traces, generated analysis summaries, reliability statistics, belief-surprise values, semantic-audit labels, held-out labels, faceted scorer outputs, run configurations, retry logs, and scoring scripts.