TRIAGE-PIEVO: EVIDENCE ADMISSION FOR PRINCIPLE-EVOLVING SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based AI Scientists automate increasingly large parts of scientific discovery, but noisy observations make their revisions unreliable. Existing principle-evolving agents act on every anomaly, whether its source is a new mechanism or an evaluation artifact; invalid revisions then bias evolution. To address this, we formulate revision as *evidence admission*: anomalies steer search immediately, but only independently verified evidence may rewrite persistent principles, and contradicted principles are retracted. We realize this in **TRIAGE-PIEVO**, which pairs an anomaly-responsive search portfolio with a staged verification gate that controls what enters principle state. Across four scientific tasks with three backbones, TRIAGE-PIEVO attains the highest average solution quality, improving over the strongest baseline of each backbone by $+7.5\% \sim 11.8\%$; on Qwen3-235B it matches that baseline's full-budget quality with one-third of the evaluation budget. In constructed stress tests, it admits none of the corrupted, hidden-source, or exploit evidence that ungated variants admit, while missing no planted mechanism. Plugged unchanged into three external research harnesses, it cuts false evidence admissions by over 60% and raises verified scores by $31\% \sim 43\%$.

1 INTRODUCTION

Autonomous scientific agents now automate scientific research within a single closed loop (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Gottweis et al., 2025). On the premise that true knowledge often lies outside the initial hypothesis space, principle-evolution agents (Pu et al., 2026; 2025) make every anomaly a decision in this loop: guide hypothesis search, or rewrite the persisted theory and explanation.

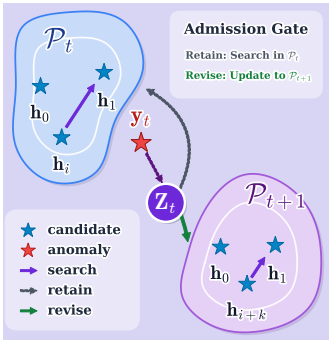


Figure 1: **Source-aware triage.** Latent source Z_t mediates the verified retain-or-revise decision.

While surprise-driven revision (Pu et al., 2026) lets agents escape their initial hypothesis space, it treats the observation's reliability as given. The revision may then degrade, because the signal it rests on cannot tell a false revision from a missed theory. Existing methods do not cover this gap: verification loops critique transient responses and leave persisted theory unexamined (Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024; Dhuliawala et al., 2024), and self-correction without external evidence is unreliable (Huang et al., 2024; Kamoi et al., 2024). Accordingly, community evaluations find that auto-research agents produce extensive artifacts yet seldom revise their internal theory under contradictory evidence (Ríos-García et al., 2026; Kirgis et al., 2026). As a result, *anomalies are read as evidence of science: corrupted feedback enters trusted scientific state*.

These limitations reduce to three challenges: (a) one anomaly score has **source ambiguity** that cannot tell a new mechanism from corrupted or misspecified evaluations; (b) the **role conflation** of one history serves both search and principle acceptance, so unreliable observations can rewrite trusted principles; and (c) the **revision irreversibility** that accepted principles cannot be retracted under contradictory evidence.

To bridge this gap, we propose **TRIAGE**, an evidence-admission layer that decouples anomaly detection from principle revision (Pu et al., 2026). TRIAGE treats each anomaly as source-ambiguous

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and routes evidence along two separately logged paths: (a) observations inform candidate selection, while (b) verification records update per-channel reliability estimates. A principle revision is trusted only after it survives independent tests on unseen cases, as shown in Figure 1. We instantiate this in **TRIAGE-PIEVO**, which pairs TRIAGE with Active Flow Expansion (De Santi et al., 2026): the portfolio allocates the evaluation budget across three complementary search roles, while the evidence-admission layer decides which evidence alters principle state.

We evaluate TRIAGE-PIEVO on four discovery tasks under three backbone LLMs. Empirical results demonstrate that **TRIAGE-PIEVO**: (a) **attains the highest performance**, raising average solution quality from PIEVO's 62.87–70.35% to 79.63–83.37%; (b) **admits no corrupted or exploit-region observations** in constructed stress tests where gate-off variants admit every one, under evaluator corruption and proxy exploitation alike (the latter further raising held-out score by 0.209); (c) **cuts false expansions** on injected noise, surrogate failure, and non-generalizing local artifacts from rate of 1.0 to 0.0 in the same tests without missing any planted mechanism, and backtracks to the correct state after mechanism reversal; and (d) **transfers unchanged to three external research harnesses**, cutting false evidence admissions by over 60% and raising verified scores by $31\% \sim 43\%$.

Our contributions are:

(a) We introduce **Evidence Admission** as the decision layer of anomaly-driven agents for changing scientific state. Anomalies are treated as source-ambiguous, and revision is governed by asymmetric costs of false revision versus missed mechanism.
(b) We propose **TRIAGE-PIEVO**, which realizes this layer through per-channel trust estimation, a principle lifecycle gated by independent held-out support.
(c) We conduct a **controlled and cross-harness evaluation** across four scientific tasks and three backbones: TRIAGE-PIEVO raises average solution quality by $+7.5\% \sim 11.8\%$ over the strongest baseline per backbone, and the corruption, hidden-source, and proxy-exploitation studies show that the gate rejects all unreliable evidence injected in these studies while missing no planted mechanism.

2 RELATED WORK

Autonomous scientific agents and evolving state. Scientific agents automate growing shares of the research cycle, spanning end-to-end paper production and multi-stage pipelines (Lu et al., 2024; Yamada et al., 2025; Tang et al., 2026; Schmidgall et al., 2025; Gottweis et al., 2025) as well as tool-coupled domain execution and closed-loop laboratory discovery (Bran et al., 2023; Boiko et al., 2023; Ruan et al., 2024; Yang et al., 2025; Ghafarollahi & Buehler, 2025; Ding et al., 2025; Ghareeb et al., 2026; Narayanan et al., 2024; Laurent et al., 2024; Mitchener et al., 2025a; Skarlinski et al., 2024; Mitchener et al., 2025b). Between iterations, these systems persist an artifact; they do not maintain a revisable scientific account. A narrower line makes the internal state editable: PIEVO and PiFlow expand a principle space on conflicting observations (Pu et al., 2026; 2025), ATLAS selects mechanism-discriminating experiments (Éltető et al., 2026), self-revising frameworks log regime changes (Wang & Buehler, 2026), and evolving architectures such as EvoMaster (Zhu et al., 2026) and ADMET-EvO (Zhou et al., 2026) pursue cumulative adaptation across tasks, with agentic laboratory world models unifying hypothesis generation and state tracking (Fu et al., 2026). But editable state is also exposed: without explicit evidence admission, anomalous or corrupted feedback rewrites persistent principles (Ríos-García et al., 2026; Kirgis et al., 2026). TRIAGE separates the history that search acts on from the evidence authorized to alter scientific state.

Active experimentation under heterogeneous sources. In active learning and Bayesian optimization, one line decides where to spend evaluations (MacKay, 1992; Houlsby et al., 2011), constraining exploration for safety (Sui et al., 2015) or scalability (Eriksson et al., 2019). Another exploits heterogeneous costs, fidelities, and failure rates across sources (Swersky et al., 2013; Kandasamy et al., 2017; Poloczek et al., 2017; Ghoreishi & Allaire, 2019), extending to unreliable channels (Mikkola et al., 2023; Fan et al., 2024), quality-controlled loops (Harris et al., 2025), and out-of-distribution expansion (De Santi et al., 2026). Both decide which observation to acquire *before* it is run, so an anomalous outcome inherits whatever reliability the acquisition rule assumed. The difference is in what the decision protects: multi-source methods weight a noisy source inside a regression model, while the object at risk here is a persistent principle that a single admitted artifact can corrupt for the

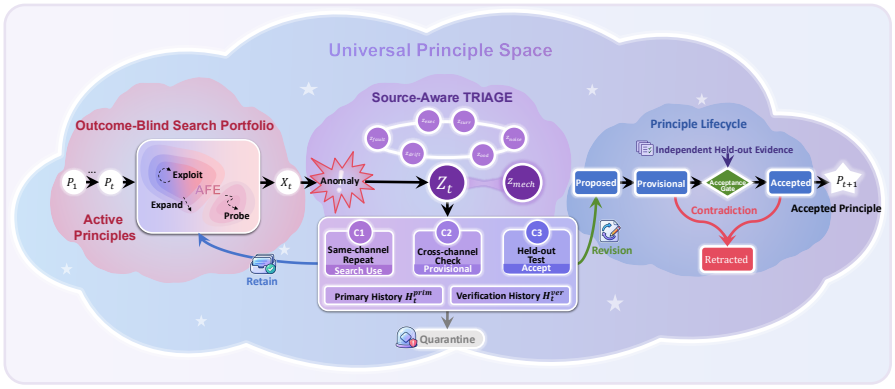


Figure 2: **Overview of TRIAGE-PIEVO.** *Left:* an outcome-blind AFE portfolio cycles candidate proposals across exploitation, frontier expansion, and mechanism-discriminating probes. *Middle:* TRIAGE treats an anomalous residual as source-ambiguous and routes it to quarantine, search history, or verification without attributing it to a source at runtime. *Right:* principles advance PROPOSED → PROVISIONAL → ACCEPTED only with independent held-out support; contradictory evidence retracts them.

rest of the run. Our AFE portfolio takes the acquisition-side view for candidate generation; TRIAGE makes the post-observation decision of whether an anomaly may update that principle.

Verification, provenance, and agent monitoring. In agent systems, verification targets transient outputs through feedback recycling (Yao et al., 2022; Shinn et al., 2023; Madaan et al., 2023), tool-based checking (Gou et al., 2024), independent verification questions (Dhuliawala et al., 2024), and evidence accumulation (Zhang et al., 2026), though self-correction without external evidence is unreliable (Huang et al., 2024; Kamoi et al., 2024). Related work argues for monitoring agentic systems before trusting them (Boston et al., 2026) and for immutable provenance trails over agent-generated artifacts (Liu et al., 2026), while Sibyl (Rangelov, 2026) uses temporal backtesting as held-out evidence. TRIAGE carries these provenance and monitoring ideas to persistent principles: same-source agreement keeps observations search-active, acceptance requires independent held-out support, and contradictory held-out evidence forces retraction.

3 METHODOLOGY

Overview. TRIAGE-PIEVO focuses on maintaining a revisable account of what observations mean. Each experiment has two possible consequences: (a) guiding the next experiment, and (b) changing the principles used to interpret future experiments. Our design follows from one premise:

> *A surprising observation can steer hypothesizing long before it is reliable enough to revise persistent scientific principles.*

As shown in Figure 2, TRIAGE-PIEVO separates *using an observation* from *believing its explanation*: it keeps exploration responsive to new signals while requiring independent evidence before revising a principle. TRIAGE-PIEVO builds on the principle-hypothesis-testing cycle of PiEvo (Pu et al., 2026).

This separation is implemented as a feedback loop with three components: (a) an evidence-admission layer checks whether an unexpected outcome generalizes to held-out cases (Section 3.2), (b) a search policy proposes experiments from outcomes that have been admitted for exploration (Section 3.3), and (c) a principle manager then updates the principle state accordingly (Section 3.4). Two histories keep these roles apart: the primary history accumulates observations the search may act on, while the verification history records linked checks that can justify (or overturn) a principle revision.

3.1 EVIDENCE ADMISSION UNDER SOURCE AMBIGUITY

Round state and the anomaly event. The agent interacts with an environment over rounds $t = 1, \ldots, T$, and each round produces exactly one record through three steps: (i) the agent proposes a testable candidate $x_t \in \mathcal{X}$; (ii) an internal predictor (the *Gaussian Process (GP) experts* $\{\mathcal{M}_P\}_{P \in \mathcal{P}_t}$ of PIEVO (Pu et al., 2026), one likelihood model per principle record, fit on the history of admitted

observations) issues a prediction $\hat{y}_t$ with predictive scale $\sigma_t > 0$; (iii) an observation channel $c_t \in \mathcal{C}$ (e.g., an evaluator or data source) evaluates x_t and returns an outcome y_t with evaluator's metadata m_t (e.g., configuration). The tuple $(x_t, \hat{y}_t, y_t, c_t, m_t)$ is the record; the deviation of y_t from $\hat{y}_t$ decides whether the record enters the primary history or opens a diagnostic state, per Definition 3.1.

Definition 3.1 (Anomaly score and anomaly event)**.** *The **anomaly score** of a record is its standardized residual $s_t := |y_t - \hat{y}_t|/\sigma_t$. Given a threshold τ_{anom} fixed before evaluation, the record is ordinary if $s_t \le \tau_{\text{anom}}$ and an anomaly if $s_t > \tau_{\text{anom}}$.*

An ordinary record is admitted directly to the primary history and triggers no further action. An anomaly, by contrast, opens a diagnostic state: it is withheld from search scores and principle updates until the verification procedure of Section 3.2 resolves it.

The anomaly score does not identify its source. A large residual is ambiguous in origin. We make this explicit by attaching to each evaluation a source variable Z_t taking values in

$$\mathcal{Z} = \underbrace{\{z_{\text{noise}}, z_{\text{fault}}, z_{\text{exec}}, z_{\text{surr}}, z_{\text{ood}}, z_{\text{drift}}\}}_{\text{artifact sources } \mathcal{Z}_{\text{art}}} \cup \underbrace{\{z_{\text{mech}}\}}_{\text{new mechanism}}, \tag{1}$$

denoting measurement noise, device fault, execution failure, mismatch of the external surrogate evaluator, out-of-distribution input, distribution drift, and a genuine new mechanism. Exactly one source generates each evaluation, and only z_{mech} justifies revision; the remaining *artifact sources* $\mathcal{Z}_{\text{art}}$ should leave the principle state untouched. Because s_t is a deterministic function of $(y_t, \hat{y}_t, \sigma_t)$ and not of Z_t, the anomaly score is uninformative about which source produced it; thresholding s_t alone therefore cannot separate a discovery from an artifact. The consequences of deciding wrongly are asymmetric, as Example 3.2 illustrates.

Example 3.2 (Two anomalies with identical scores)**.** *Suppose two rounds produce the same large score s_t. In round t_1, a corrupted evaluator channel (z_{fault}) inflates the outcome of a mediocre candidate; revising $\mathcal{P}_t$ on this record redirects every subsequent prediction toward an artifact, corrupting all future rounds. In round t_2, a novel candidate (z_{mech}) violates the current principles; retaining $\mathcal{P}_t$ postpones the discovery until stronger evidence arrives. The first error propagates through persistent state; the second only costs time.*

To address this asymmetry, we define the admission decision as choosing $d \in \{\mathsf{retain}, \mathsf{revise}\}$ after observing an anomaly, with loss

$$L(d, z) = \ell_{\text{FR}} \mathbb{1}[d = \mathsf{revise}, z \in \mathcal{Z}_{\text{art}}] + \ell_{\text{FM}} \mathbb{1}[d = \mathsf{retain}, z = z_{\text{mech}}], \tag{2}$$

where ℓ_{FR} and ℓ_{FM} price false revision and false retention, respectively, with $\ell_{\text{FR}} \gg \ell_{\text{FM}}$ per Example 3.2. Under this loss, the risk-optimal action is to revise only when the posterior $q(z_{\text{mech}} \mid y_t, \hat{y}_t, m_t)$ exceeds $\ell_{\text{FR}}/(\ell_{\text{FR}} + \ell_{\text{FM}})$, a threshold close to one.

Remark. The *reference model* of Eq. (2) implies three requirements: (i) anomaly detection must trigger further evidence collection; (ii) the evidence required for revision must be independent of the channel c_t; and (iii) revising persistent state must demand strictly stronger evidence than search use does.

3.2 Escalating from Repeats to Held-Out Probes

Because an anomaly may affect candidate search, principle proposal, or principle revision, TRIAGE applies three checks (C1∼C3) of increasing independence from the originating experiment channel c_t, stopping at the first failure. Each check rules out one class of artifact sources in $\mathcal{Z}_{\text{art}}$ (see Eq. (1)) and grants a single privilege.

C1: Rules out transient faults; unlocks search use. Re-evaluating x_t on c_t yields $v_t^{c_t}$. Agreement, $|y_t - v_t^{c_t}| \le \epsilon_{\text{rep}}$, excludes transient noise and execution failures ($z_{\text{noise}}, z_{\text{exec}}$). This cannot authorize revision, since repeating on c_t replicates whatever systematic bias c_t carries.

C2: Rules out channel-specific corruption; unlocks revision proposals. Re-evaluating x_t on an independent channel $c' \neq c_t$ yields $v_t^{c'}$. Agreement, $|v_t^{c_t} - v_t^{c'}| \le \epsilon_{\text{ind}}$, excludes corruption specific to c_t (z_{fault}), so the anomaly may propose a principle and become provisional (see Section 3.4).

C3: Rules out non-generalizing artifacts; unlocks acceptance. Evaluation by an independent evaluator on held-out inputs excludes effects tied to the original evaluation context ($z_{\text{surr}}, z_{\text{ood}}, z_{\text{drift}}$), and is the evidence the acceptance gate counts (see Section 3.4). A record failing any check is quarantined, i.e., excluded from both histories' influence until resolved; a verified replacement may remain search-active but cannot update persistent state by itself.

Reconciliation outcomes update per-channel reliability, so that frequently contradicted channels are progressively excluded from verification duty. Each channel c carries a reliability estimate $r_c \sim \text{Beta}(\alpha_c, \beta_c)$ with trust score $p_c^{\text{trust}} = \alpha_c/(\alpha_c + \beta_c)$; agreement increments α_c and disagreement increments β_c for every participating channel (see Algorithm 1).

Remark. Trust selects which channels may verify, but does not lower the standard of evidence. All thresholds are fixed before evaluation and are independent of p_c^{trust}, so no amount of accumulated trust lets a channel certify its own anomalies.

3.3 Proposing Candidates without Seeing the Gate

We use Active Flow Expansion (AFE) (De Santi et al., 2026) as a search component in TRIAGE-PIEVO . It proposes candidates in *three roles*: best-region **exploitation**, frontier **expansion**, and mechanism-discriminating **probes**.

Scoring candidates. For each role r, AFE scores a candidate $x \in \mathcal{X}$ by four terms computed from the GP experts of Section 3.1, refit on the admitted primary history:

$$A_r(x) = w_r^{\mu}\,\hat{\mu}(x) + w_r^{\sigma}\,\hat{\sigma}(x) + w_r^{n}\,N(x) + w_r^{d}\,D(x),$$

where $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are the GP experts' predicted outcome and predictive uncertainty, $N(x)$ is the distance of x from already evaluated inputs, and $D(x)$ is a descriptor-novelty measure. The role weights $\{w_r^{\mu}, w_r^{\sigma}, w_r^{n}, w_r^{d}\}$ are fixed a priori and differ across roles: exploitation weights $\hat{\mu}$, expansion weights N and D, and probes weight $\hat{\sigma}$ in regions where competing principles disagree (Figure 3).

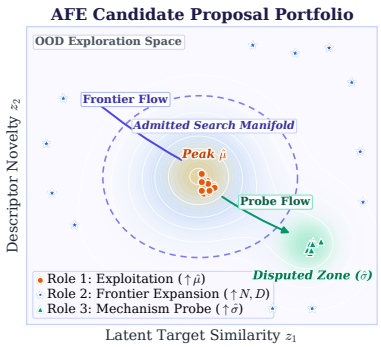


Figure 3: **AFE portfolio schematic.** Three complementary roles on a shared design manifold: exploitation ($\uparrow \hat{\mu}$), frontier expansion ($\uparrow N, D$), and mechanism probes ($\uparrow \hat{\sigma}$).

Information boundary between AFE and TRIAGE . Two append-only histories keep search evidence and revision evidence apart: the *primary history* H_t^{prim} stores the records the search policy may act on, and the *verification history* H_t^{ver} stores the C1–C3 checks (see Section 3.2), each linked to the primary record it concerns. "Outcome-blind" is then a one-way boundary over these histories: AFE reads only H_t^{prim} and candidate descriptors, outputs x_t, and is forbidden from verification channels, quarantine flags, and gate state; TRIAGE writes back, entering the admitted, substituted, or quarantined outcome into H_{t+1}^{prim} for the next refit of the GP experts (PIEVO 's per-principle likelihood models $p(y \mid h, P)$ (Pu et al., 2026), which we use unmodified). Unable to observe gate state, the search policy cannot route unverified evidence into persistent state; the acceptance gate of Section 3.4 is the sole write path to principle state.

3.4 A Gated Principle Lifecycle with Reversible Revision

The final component determines how verified evidence changes persistent state. Each principle P is a record $(\text{parent}(P), E(P), \text{state}(P))$: a pointer to the principle it refines, its linked evidence set, and a lifecycle state that advances through PROPOSED $\longrightarrow$ PROVISIONAL $\longrightarrow$ ACCEPTED, with any state sent to RETRACTED upon contradiction. We specify the two forward transitions and the retraction rule in turn.

Acceptance requires admissible evidence. An anomalous record passing C1–C2 (see Section 3.2) creates or updates a PROPOSED principle, and C2's cross-channel agreement promotes it to PROVISIONAL. Acceptance, however, is gated by the following class of evidence:

Definition 3.3 (Admissible evidence for revision)**.** *A record e is* admissible *for principle P if (i) it is produced by an evaluator process sharing no memory, cache, or random seed with the*

*generator (*independence*), and (ii) it is measured on X_{heldout} disjoint from the generator-visible X_{train} (*held-out*). An admissible record* supports *P if its held-out gain $\Delta_{\text{heldout}}(e; P) \geq \tau_{\text{gen}}$ and* contradicts *P if $\Delta_{\text{heldout}}(e; P) \leq -\tau_{\text{gen}}$; records in between are uninformative.*

Writing n_P^+ and n_P^- for the numbers of supporting and contradicting admissible records in $E(P)$, PROVISIONAL $\longrightarrow$ ACCEPTED $\Longleftrightarrow n_P^- = 0 \ \wedge \ n_P^+ \geq k$. The two conditions of Definition 3.3 each remove one failure mode: independence prevents corroboration from being manufactured by the proposing channel, and the held-out requirement makes acceptance test generalization rather than fit.

Standing retraction of contradicted principles. A single contradicting admissible record ($n_P^- \geq 1$) moves any active principle (PROPOSED, PROVISIONAL, or ACCEPTED) to RETRACTED, and the retraction propagates through $\text{parent}(\cdot)$, so descendants of a contradicted principle cannot remain trusted. Retraction is standing: every newly admitted record is scored against each active principle it bears on, so evidence arriving after acceptance can still overturn an ACCEPTED principle (Algorithm 1).

The full loop of TRIAGE-PIEVO appears in Algorithm 1; thresholds and sensitivity analyses are in Appendix A.

4 EXPERIMENTS

We evaluate TRIAGE-PIEVO along four axes: (1) matched downstream performance across four scientific discovery tasks under three backbone LLMs; (2) search efficiency in optimization speed, candidate coverage, and computational cost; (3) controlled stress tests of the admission gate under evaluator corruption, hidden anomaly sources, mechanism reversal, proxy exploitation, and cross-harness deployment; and (4) ablations isolating search expansion from evidence admission.

4.1 EXPERIMENT SETUP

Tasks and Benchmarks. Following Pu et al. (2026), we evaluate TRIAGE-PIEVO on four optimization tasks: Molecular Bio-activity (MBO), Nanohelix (NHO), Superconductor (SPO), and Transition Metal Complex (TMC). Each task is scored by its domain evaluator and maximizes a task-specific objective under a shared budget of 24 evaluations.

Baselines. We evaluate three categories of scientific agents: (a) **generic and classic**: Vanilla MAS and ReAct (Yao et al., 2022), which represent tool-use loops without persistent scientific state; (b) **autonomous research**: AI Scientist v1 (Lu et al., 2024), AI Scientist v2 (Yamada et al., 2025), and AI Researcher (Tang et al., 2026), which automate multi-stage pipelines over transient artifacts; and (c) **principle-evolving**: PiFlow (Pu et al., 2025) and PIEVO (Pu et al., 2026), the state of the art in revising a principle set on anomalous observations. Our method places exactly this response under admission control.

Implementation Details. We evaluate three backbones: Qwen3-235B (Team, 2025), Gemini-2.5-Flash-Lite (Google, 2025), GPT-5.6-Luna (OpenAI, 2026) without extended chain-of-thought, over three fixed seeds, reported as mean $\pm$ std. Gate parameters in TRIAGE-PIEVO are fixed a priori across all tasks and backbones. Full specification and sensitivity sweeps in Appendix A.

4.2 EVALUATION METRICS

With $y_{1:T}$ the primary outcomes of a run, we employ metrics that capture abilities to (a) reach high-quality solutions, (b) reach them early, and (c) cover the candidate space broadly:

a. Solution Quality (SQ, %, ↑). The best outcome found, normalized by a published per-task reference μ_{ref} from PIEVO (Pu et al., 2026): $\text{SQ} = 100 \cdot {}^{\max_t y_t}/_{\mu_{\text{ref}}}$, with $\mu_{\text{ref}} = 6.5$ (MBO), 2.0 (NHO), 298.5 (SPO), and 493.8 (TMC). These references are not attainable upper bounds, so SQ passes 100 when a run beats the reference; we apply no clipping.

b. Area Under the Optimization Curve (AUOC, %, ↑). The budget-averaged best-so-far quality, $\text{AUOC} = 100 \cdot \frac{1}{T} \sum_t {}^{\max_{t' \leq t} y_{t'}}/_{\mu_{\text{ref}}}$, crediting how early quality is reached.

Table 1: **Solution Quality (SQ, %) ↑ across four discovery tasks and three backbone LLMs.** Entries report mean ± std over three seeds; **Bold** indicates the best result within each backbone.

Model / Method	MBO	NHO	SPO	TMC	Average
Qwen3-235B (No Thinking)					
Vanilla MAS	83.35 ± 29.95	52.83 ± 20.36	37.29 ± 0.75	56.34 ± 3.69	57.45
ReAct	75.48 ± 4.94 ↓ 0.9x	61.95 ± 17.08 ↑ 1.2x	37.49 ± 0.46	72.80 ± 4.48 ↑ 1.3x	61.93 ↑ 1.1x
AI Scientist v1	73.37 ± 39.82 ↓ 0.9x	41.50 ± 3.12 ↓ 0.8x	37.04 ± 0.61	63.96 ± 8.93 ↑ 1.1x	53.97 ↓ 0.9x
AI Scientist v2	84.40 ± 20.07	52.34 ± 15.56	41.40 ± 1.15 ↑ 1.1x	63.52 ± 2.86 ↑ 1.1x	60.42 ↑ 1.1x
AI Researcher	105.70 ± 21.92 ↑ 1.3x	37.21 ± 5.68 ↓ 0.7x	37.39 ± 0.85	78.80 ± 3.76 ↑ 1.4x	64.78 ↑ 1.1x
PiFlow	81.89 ± 30.49	56.15 ± 15.23 ↑ 1.1x	37.21 ± 0.52	54.58 ± 2.70	57.46
PIEVO	97.84 ± 27.49 ↑ 1.2x	65.53 ± 21.90 ↑ 1.2x	37.71 ± 0.53	80.30 ± 4.57 ↑ 1.4x	70.35 ↑ 1.2x
Triage-PiEvo (ours)	**132.00** ± 9.84 ↑ 1.6x	**72.89** ± 17.80 ↑ 1.4x	**42.00** ± 1.20 ↑ 1.1x	**81.50** ± 4.02 ↑ 1.4x	**82.10** ↑ 1.4x
Gemini-2.5-Flash-Lite (No Thinking)					
Vanilla MAS	62.06 ± 14.09	71.98 ± 1.01	36.95 ± 0.46	66.64 ± 17.42	59.41
ReAct	58.83 ± 6.44 ↓ 0.9x	69.90 ± 1.50	32.06 ± 5.75 ↓ 0.9x	56.31 ± 2.37 ↓ 0.8x	54.27 ↓ 0.9x
AI Scientist v1	75.60 ± 22.03 ↑ 1.2x	73.58 ± 2.23	37.86 ± 2.55	55.70 ± 2.62 ↓ 0.8x	60.68
AI Scientist v2	99.30 ± 22.73 ↑ 1.6x	74.87 ± 2.00	30.74 ± 3.96 ↓ 0.8x	54.80 ± 4.09 ↓ 0.8x	64.93 ↑ 1.1x
AI Researcher	66.47 ± 16.70 ↑ 1.1x	65.39 ± 9.16 ↓ 0.9x	31.62 ± 3.82 ↓ 0.9x	63.46 ± 13.85	56.74
PiFlow	96.59 ± 11.71 ↑ 1.6x	74.24 ± 4.06	34.91 ± 3.66 ↓ 0.9x	68.22 ± 8.71	68.49 ↑ 1.2x
PIEVO	72.59 ± 6.70 ↑ 1.2x	81.00 ± 3.06 ↑ 1.1x	38.02 ± 0.50	78.50 ± 7.60 ↑ 1.2x	67.53 ↑ 1.1x
Triage-PiEvo (ours)	**115.00** ± 10.00 ↑ 1.9x	**83.00** ± 5.00 ↑ 1.2x	**40.50** ± 4.00 ↑ 1.1x	**80.00** ± 4.00 ↑ 1.2x	**79.63** ↑ 1.3x
GPT-5.6-Luna (No Thinking)					
Vanilla MAS	91.84 ± 18.08	66.26 ± 7.14	36.42 ± 0.50	68.07 ± 3.91	65.65
ReAct	132.00 ± 22.98 ↑ 1.4x	69.98 ± 0.14 ↑ 1.1x	36.22 ± 0.34	65.37 ± 5.39	75.89 ↑ 1.2x
AI Scientist v1	91.16 ± 19.35	53.02 ± 12.32 ↓ 0.8x	35.52 ± 4.90	62.44 ± 6.24 ↓ 0.9x	60.54 ↓ 0.9x
AI Scientist v2	89.63 ± 31.45	65.13 ± 15.46	36.77 ± 5.05	58.29 ± 0.95 ↓ 0.9x	62.46
AI Researcher	77.86 ± 15.38 ↓ 0.8x	73.92 ± 9.01 ↑ 1.1x	36.28 ± 0.24	72.93 ± 2.06 ↑ 1.1x	65.25
PiFlow	128.77 ± 18.74 ↑ 1.4x	46.62 ± 17.77 ↓ 0.7x	31.49 ± 5.94 ↓ 0.9x	59.59 ± 5.23 ↓ 0.9x	66.62
PIEVO	69.75 ± 6.20 ↓ 0.8x	67.82 ± 10.87	37.41 ± 0.53	76.50 ± 1.78 ↑ 1.1x	62.87
Triage-PiEvo (ours)	**136.00** ± 9.84 ↑ 1.5x	**78.18** ± 5.44 ↑ 1.2x	**40.80** ± 5.70 ↑ 1.1x	**78.50** ± 3.75 ↑ 1.2x	**83.37** ↑ 1.3x

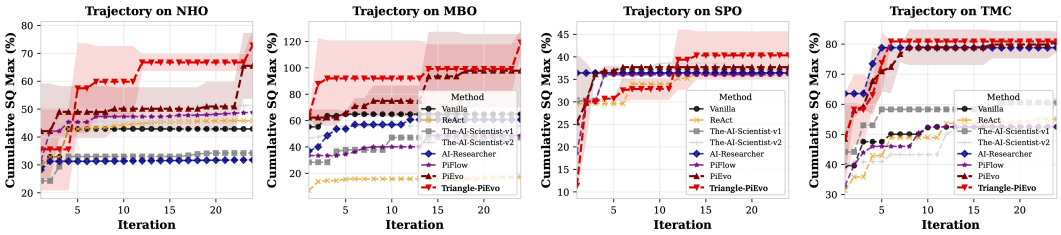


Figure 4: **Best-so-far trajectories across four benchmarks with Qwen3-235B backbone.** Across all scientific domains, TRIAGE-PIEVO accumulates high-quality candidates and sustains the highest solution quality.

c. Average Pairwise Distance (APD, ↑). The mean pairwise cosine distance between proposed candidates. A higher APD indicates broader candidate coverage.

4.3 PERFORMANCE AND EFFICIENCY ANALYSIS

Table 1 compares TRIAGE-PIEVO against seven baselines on solution quality (SQ) across four tasks and three backbones, and Table 2 reports the corresponding coverage (APD) and accumulation speed (AUOC).

Obs.❶: TRIAGE-PIEVO attains the highest average SQ under every backbone. TRIAGE-PIEVO attains the highest SQ in every task–backbone and the highest average under every backbone, improving on the strongest baseline of each backbone by 11.75%, 11.14%, and 7.48% (Table 1); because the gain appears under all three backbones, the failure mode that evidence admission repairs is a property of the agent class, not of any single model.

Obs.❷: TRIAGE-PIEVO leads on average coverage and on accumulation speed. TRIAGE-PIEVO achieves a 1.16×–1.22× faster accumulation (Table 2), while baselines trade the two facets off (e.g., AI Researcher (Tang et al., 2026) reaches 89.21 AUOC on MBO but at the lowest average coverage); on NHO, ReAct and AI Scientist v2 attain higher coverage but lower AUOC, so the lead

Table 2: **Average Pairwise Distance (APD, ↑) and Area Under the Optimisation Curve (AUOC, ↑) across four tasks and three backbone LLMs.** Higher APD indicates broader candidate coverage, while higher AUOC indicates faster accumulation of high-quality candidates.

Model / Method	MBO		NHO		SPO		TMC		Average	
	APD ↑	AUOC ↑	APD ↑	AUOC ↑	APD ↑	AUOC ↑	APD ↑	AUOC ↑	Avg APD ↑	Avg AUOC ↑
Qwen3-235B (No Thinking)										
Vanilla MAS	53.58±15.25	70.09±17.75	32.23±3.72	49.33±17.82	52.03±1.05	36.34±0.38	51.89±3.85	52.67±3.49	47.43	52.11
ReAct	59.43±8.60	70.58±7.95	**33.31**±1.86	47.46±7.04	56.13±4.49	35.74±0.76	72.93±0.33	66.95±4.67	55.45 ↑ 1.2x	55.18 ↑ 1.1x
AI Scientist v1	52.93±15.22	56.16±26.05	32.98±2.32	38.84±2.06	42.55±5.61	35.62±1.46	64.43±7.02	57.50±6.15	48.22	47.03 ↓ 0.9x
AI Scientist v2	60.26±10.57	76.03±7.92	28.39±3.49	39.51±3.29	48.18±1.41	37.03±0.22	55.66±2.28	54.75±6.66	48.12	51.83
AI Researcher	59.17±3.77	89.21±19.08	25.51±3.45	34.51±3.80	47.62±3.55	36.82±0.41	52.46±6.03	74.19±0.70	46.19	58.68 ↑ 1.1x
PiFlow	64.99±2.39	64.04±14.87	28.07±1.85	52.93±15.82	51.24±4.14	35.90±1.33	57.76±3.35	50.26±1.85	50.52 ↑ 1.1x	50.78
PIEVO	63.41±10.78	84.14±16.22	23.80±5.23	52.50±13.36	54.41±1.42	36.83±0.86	71.76±2.58	74.93±5.05	53.35 ↑ 1.1x	62.10 ↑ 1.2x
Triage-PiEvo (ours)	**73.71**±2.86	**110.00**±18.41	23.67±0.79	**61.21**±11.52	**57.34**±0.26	**38.50**±1.80	**74.20**±1.41	**77.20**±2.95	**57.23** ↑ 1.2x	**71.73** ↑ 1.4x
Gemini-2.5-Flash-Lite (No Thinking)										
Vanilla MAS	65.06±5.46	46.01±15.24	33.55±0.90	61.71±2.41	53.65±1.10	35.45±0.05	57.23±5.29	62.88±15.97	52.37	51.51
ReAct	62.08±1.15	44.54±2.54	34.46±0.57	58.08±5.33	42.17±12.13	25.95±8.19	52.75±5.72	52.03±0.12	47.87 ↓ 0.9x	45.15 ↓ 0.9x
AI Scientist v1	52.02±15.45	58.73±5.53	33.13±0.89	64.99±5.57	47.13±4.98	33.96±1.93	56.53±1.71	51.60±0.09	47.20 ↓ 0.9x	52.32
AI Scientist v2	46.27±12.65	59.35±11.81	**35.09**±0.38	67.18±1.82	43.28±0.12	24.33±7.17	42.16±9.83	52.77±3.83	41.70 ↓ 0.8x	50.91
AI Researcher	60.38±7.64	45.98±15.08	33.16±3.51	58.94±6.79	41.65±9.04	28.78±2.15	55.97±2.51	59.77±14.66	47.79 ↓ 0.9x	48.37 ↓ 0.9x
PiFlow	62.64±6.57	68.05±8.07	29.73±4.14	65.72±5.25	44.03±2.95	33.27±2.48	66.17±6.06	63.89±6.66	50.64	57.73 ↑ 1.1x
PIEVO	64.40±1.50	71.40±7.88	23.38±0.37	65.97±3.31	53.88±0.59	35.02±1.92	70.35±0.76	69.74±7.30	53.00	60.53 ↑ 1.2x
Triage-PiEvo (ours)	**70.00**±5.00	**108.00**±12.00	24.00±2.00	**76.00**±8.00	**57.00**±1.50	**36.00**±4.00	**72.50**±2.50	**75.00**±6.00	**55.88** ↑ 1.1x	**73.75** ↑ 1.4x
GPT-5.6-Luna (No Thinking)										
Vanilla MAS	64.66±4.12	69.17±12.39	27.30±1.90	60.18±3.66	48.73±2.51	34.43±0.60	71.55±1.07	63.26±4.82	53.06	56.76
ReAct	38.36±20.70	86.71±13.76	**31.58**±2.81	59.57±2.38	49.26±2.90	33.88±0.86	66.05±7.21	62.28±5.12	46.31 ↓ 0.9x	60.61 ↑ 1.1x
AI Scientist v1	55.57±18.24	71.58±4.38	29.17±3.01	44.85±6.39	47.10±0.71	29.87±4.50	53.91±3.25	54.73±0.57	46.44 ↓ 0.9x	50.26 ↓ 0.9x
AI Scientist v2	51.59±19.99	60.12±11.92	24.07±6.68	54.21±13.33	45.21±3.80	32.19±5.66	49.34±11.99	51.65±1.98	42.55 ↓ 0.8x	49.54 ↓ 0.9x
AI Researcher	66.12±1.10	70.56±16.40	27.41±4.25	65.81±4.10	50.05±3.05	32.56±1.91	71.88±0.71	65.59±2.27	53.86	58.63
PiFlow	48.90±24.99	79.51±4.35	17.19±2.39	38.66±10.82	37.93±7.95	28.15±7.38	56.56±3.30	56.09±6.93	40.14 ↓ 0.8x	50.60 ↓ 0.9x
PIEVO	64.05±1.98	62.95±1.52	23.83±0.49	52.04±7.69	53.07±0.84	36.29±1.21	73.20±0.54	74.81±0.97	53.54	56.52
Triage-PiEvo (ours)	**73.72**±2.84	**104.00**±19.33	24.08±1.10	**68.77**±8.52	**57.87**±0.88	**38.10**±1.60	**74.80**±1.20	**76.80**±3.20	**57.62** ↑ 1.1x	**71.92** ↑ 1.3x

holds on average but not in every task–metric cell. The gate therefore filters evidence for revision without constricting the search.

Obs.❸: TRIAGE-PIEVO opens its margin early and holds it at a bounded cost. On Qwen3-235B, TRIAGE-PIEVO matches the strongest baseline's full-budget result at one-third of the budget (Figure 4; Appendix B, Figure 7). The gate adds $+13.6\%$ wall-clock time over PIEVO (134.5s vs. 118.4s per run), while the multi-agent and tree-search frameworks it outperforms cost $1.28\times$–$1.85\times$ more (AI Scientist v2: 172.4s; AI Researcher: 188.2s; PiFlow: 248.6s). Figure 8 places TRIAGE-PIEVO on the quality–cost efficiency frontier (Appendix C).

4.4 STRESS TESTS UNDER UNRELIABLE FEEDBACK

We further intervene on the feedback channel and compare gated and ungated variants across two complementary regimes: (i) *Constructed Stress Tests* (Appendix E), where corruptions, proxy exploits, and distribution shifts are synthetically injected with deterministic ground truth; and (ii) *Organic Benchmark Tasks* (§4.5, Table 9), evaluating continuous, unlabelled noise. We use AutoResearchEval (Fei et al., 2026) for evaluation; see Appendix H for details.

Obs.❹: In the constructed stress tests, TRIAGE-PIEVO rejects the unreliable evidence in every condition and loses no planted mechanism. Across evaluator corruption, hidden anomaly sources, and proxy exploitation, TRIAGE-PIEVO reduces the admission rate on unreliable evidence from the ungated variants' 1.0 to 0.0 (including exploits that reproduce on the proxy), while recovering planted mechanisms with 1.0 recall and, in the proxy-exploitation study, improving held-out score by $+0.209$ (Tables 7 and 8 in Appendix E). The organic gate ablation (§4.5 and Table 9) traces this discrimination to the held-out criterion: the same-source-repeat variant rules out repetition as an explanation. The gate also acts in reverse and transfers: it flags a mechanism reversal at t=17, retracts the contradicted principle with its descendants one round later (t=18), and recovers the new optimum by t=23, where PIEVO stagnates; attached unchanged to three external harnesses, it raises the externally verified score by 0.087–0.098 at additional token and runtime cost (Figure 5).

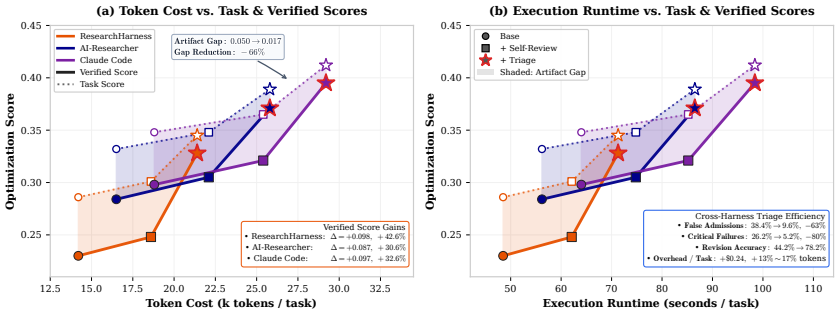


Figure 5: **Cost-efficiency trade-offs across autonomous research harnesses.** (a) Token expenditure versus reported Task Score and externally Verified Score. TRIAGE narrows the gap between reported and verified scores (shaded region) at small additional token cost. (b) Wall-clock execution runtime versus optimization performance across ResearchHarness, AI-Researcher, and Claude Code. Full numerical results are reported in Table 6 (Appendix D).

4.5 ABLATION STUDY

We ablate at three levels: the candidate-generation and evidence-admission modules of TRIAGE-PIEVO, each mechanism of the TRIAGE gate, and the gate's decision thresholds.

Module Ablations. Table 3 reports a 48-cell factorial study on Qwen3-235B at budget 8. Triage-only falls below PIEVO on SQ (60.60% vs. 64.78%): without search expansion, filtering only narrows an already narrow candidate flow, and the mild, unlabelled noise of the organic tasks gives it little to filter. TRIAGE-PIEVO w/o Gate restores exploration but admits whatever the evaluator reports. TRIAGE determines what may change persistent state. The gate therefore trades a small SQ loss under routine noise (Table 3) for a large gain under the injected unreliable feedback of §4.4.

Table 3: **Module ablations across four tasks.** SQ and AUOC are averaged across tasks and seeds.

Variants	Avg SQ ↑	Avg AUOC ↑
TRIAGE-PIEVO	**65.98**	**60.45**
w/o Gate	62.73 (Δ=−3.25)	58.23 (Δ=−2.22)
Triage-only	60.60 (Δ=−5.38)	55.08 (Δ=−5.37)
PIEVO (Pu et al., 2026)	64.78 (Δ=−1.20)	58.15 (Δ=−2.30)

Gate Mechanism Ablations. Ablating each component of the admission policy on organic benchmark tasks (Table 9, Appendix F) locates the protective effect in the held-out criterion: removing held-out verification raises false revision from 4.8% to 34.5%, and treating a same-source repeat as independent raises it to 45.8%. Both admit evidence that is reproducible but not transferable, the artifact case of Obs.❹. Removing channel reliability raises false revision to 19.2% through a different route: verification is then routed without regard to which channels have been reconcilable. The gate therefore depends on both the standard of proof and the choice of who supplies it. Mechanism recall here (92.4%) sits below the 100% of the constructed tests (Table 7): planted mechanisms carry effect sizes $\Delta \geq 2.5\sigma_0$ that clear τ_{gen} throughout the constructed tests, whereas organic tasks also present weaker effects that can fall below it.

Hyperparameter Sensitivity. Sweeps over the gate thresholds (Appendix A, Figure 6) show the fixed operating point is not finely tuned: the default $\tau_{\text{gen}} = 0.16$ lies inside the zero-error region, tolerating over ∼80% held-out signal attenuation before a true mechanism is rejected, and the default $\tau_{\text{anom}} = 0.85$ sits on a flat zero-miss plateau before coverage collapses at more restrictive thresholds.

5 CONCLUSION

An anomalous residual is source-ambiguous: new mechanisms and evaluation artifacts produce the same score. TRIAGE separates the two uses of such evidence: observations inform search at once, while persistent principles change only on independent held-out support and are retracted on held-out contradiction. Instantiated in TRIAGE-PIEVO, this separation improves solution quality on all four tasks and three backbones and eliminates false admissions in every controlled corruption study, at a +13.6% runtime overhead. The gate's guarantees are established for the injected artifact classes of §4.4; whether organic evaluator drift in deployed pipelines triggers the same staged escalation remains open.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and presentation of the manuscript. All AI-assisted content was carefully reviewed and verified by the authors. All scientific ideas, methodological designs, experimental results, and conclusions were developed and validated by the authors. We take full responsibility for the final content of this work.

REFERENCES

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Marisa Ferrara Boston, Glen Hanson, Effi Georgala, JD Hudgens, and Heather Frase. Monitoring agentic systems before they're reliable. *arXiv preprint arXiv:2606.02494*, 2026.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Riccardo De Santi, Bruce Lee, Cristian Perez Jensen, Kimon Protopapas, Sophia Tang, Cheng-Hao Liu, Pranam Chatterjee, Yisong Yue, and Andreas Krause. Active flow expansion for out-of-distribution discovery: from theory to molecules. *arXiv preprint arXiv:2606.08802*, 2026.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972, 2025.

Noémi Éltető, Nathaniel D Daw, Kimberly L Stachenfeld, and Kevin J Miller. Atlas: Active theory learning for automated science. *arXiv preprint arXiv:2606.12386*, 2026.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32, 2019.

Mingzhou Fan, Byung-Jun Yoon, Edward Dougherty, Francis Alexander, Nathan Urban, Raymundo Arróyave, and Xiaoning Qian. Multi-fidelity bayesian optimization with multiple information sources of input-dependent fidelity. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina Ciobanu, Qingqing Mao, and Ritankar Das. How do agents fail on autoresearch: End-to-end diagnostic evaluation on 100 real-world frontier research tasks. *arXiv preprint arXiv:2608.14905*, 2026.

Kelsey Fu, Luna Lyu, Shuzhen Li, Suozhi Huang, Suheng Xu, Joy Zheng, Xuefeng Liu, Shilong Liu, Greg Barbone, Ying Liu, et al. Agentic laboratories of the future: Towards world models for scientific discovery. *Preprints.org*, 2026.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.

Seyede Fatemeh Ghoreishi and Douglas Allaire. Multi-information source constrained bayesian optimization: Sf ghoreishi and d. allaire. *Structural and Multidisciplinary Optimization*, 59(3): 977–991, 2019.

Google. Gemini 2.5 flash-lite. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite, 2025. Accessed: 2026-09-23.

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*, 2, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Sumner B Harris, Rama Vasudevan, and Yongtao Liu. Active oversight and quality control in standard bayesian optimization for autonomous experiments. *npj Computational Materials*, 11(1):23, 2025.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity bayesian optimisation with continuous approximations. In *International conference on machine learning*, pp. 1799–1808. PMLR, 2017.

Peter Kirgis, Sayash Kapoor, Andrew Schwartz, Stephan Rabanser, David Africa, Konstantinos Voudouris, Viet Nguyen, Toby Pilditch, Magda Dubois, Harry Coppock, et al. Can ai agents conduct open-ended ai research? early evidence from two case studies. *arXiv preprint arXiv:2607.27191*, 2026.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, et al. The last human-written paper: Agent-native research artifacts. *arXiv preprint arXiv:2604.24658*, 2026.

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.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Petrus Mikkola, Julien Martinelli, Louis Filstroff, and Samuel Kaski. Multi-fidelity bayesian optimization with unreliable information sources. In *International Conference on Artificial Intelligence and Statistics*, pp. 7425–7454. PMLR, 2023.

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*, 2025a.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025b.

Siddharth Narayanan, James D Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G Rodriques, et al. Aviary: training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154*, 2024.

OpenAI. Gpt-5.6. https://openai.com/index/gpt-5-6/, 2026. Accessed: 2026-09-23.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. *Advances in neural information processing systems*, 30, 2017.

Yingming Pu, Tao Lin, and Hongyu Chen. Piflow: Principle-aware scientific discovery with multi-agent collaboration. *arXiv preprint arXiv:2505.15047*, 2025.

Yingming Pu, Tao Lin, and Hongyu Chen. Principle-evolvable scientific discovery via uncertainty minimization. *arXiv preprint arXiv:2602.06448*, 2026.

Blagoy Rangelov. Sibyl: Temporal backtesting for literature-based scientific discovery with large language model agents. In *ICML 2026 AI for Science Workshop*, 2026.

Martiño Ríos-García, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, NM Krishnan, and Kevin Maik Jablonka. Ai scientists produce results without reasoning scientifically. *arXiv preprint arXiv:2604.18805*, 2026.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications*, 15(1):10160, 2024.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pp. 997–1005. PMLR, 2015.

Kevin Swersky, Jasper Snoek, and Ryan Adams. Multi-task bayesian optimization. *Advances in neural information processing systems*, 26, 2013.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2026.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Fiona Y Wang and Markus J Buehler. Self-revising discovery systems for science: A categorical framework for agentic artificial intelligence. *arXiv preprint arXiv:2606.01444*, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *International Conference on Learning Representations*, volume 2025, pp. 33251–33277, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yichi Zhang, Nabeel Seedat, Yinpeng Dong, Peng Cui, Jun Zhu, and Mihaela van de Schaar. Guideline-grounded evidence accumulation for high-stakes agent verification. *arXiv preprint arXiv:2603.02798*, 2026.

Yiling Zhou, Yilin Wang, Jianmin Wang, Heqin Zhu, Zirui Wang, Chang-yu Hiesh, Kejun Ying, Jiaqi Wang, Yuzhi Xu, Tingjun Hou, et al. Admet-evo: a self-evolving scientific agent for sustained research across heterogeneous tasks. *arXiv preprint arXiv:2609.10121*, 2026.

Xinyu Zhu, Yuzhu Cai, Zexi Liu, Cheng Wang, Fengyang Li, Wenkai Jin, Wanxu Liu, Zehao Bing, Bingyang Zheng, Jingyi Chai, et al. Evomaster: A foundational evolving agent framework for agentic science at scale. *arXiv preprint arXiv:2604.17406*, 2026.

A Evidence-Admission Hyperparameters and Sensitivity Analysis

Table 4 summarizes the default hyperparameter values for the TRIAGE admission gate. All five parameters are fixed across all benchmark tasks and backbones without task-specific tuning.

Table 4: Evidence-admission gate hyperparameters. All values are fixed across all experiments.

Constant	Role	Value
τ_{anom}	Residual threshold to open diagnostic state	0.85
ϵ_{rep}	Same-channel agreement tolerance	0.12
ϵ_{ind}	Cross-channel agreement tolerance	0.12
τ_{gen}	Held-out gain required to authorize a revision	0.16
k	Independent held-out records required for acceptance	2

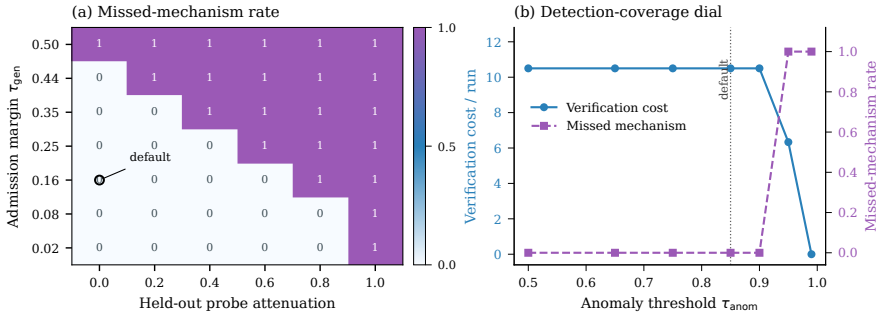


Figure 6: **Gate sensitivity analysis.** (a) Missed-mechanism rate as a joint function of the held-out gain threshold τ_{gen} and held-out probe attenuation. The default operating point (circle) lies in the zero-error region, and the boundary moves monotonically with effective held-out gain. (b) Verification cost and missed-mechanism rate as a function of the anomaly threshold τ_{anom}; the default (dashed) lies on the flat zero-miss plateau before coverage collapses at more restrictive thresholds.

Figure 6 presents the two-dimensional sensitivity phase map across decision thresholds; the default operating point and its error margin are discussed in §4.5.

B Exploration Diversity (APD) and Exploitation Trajectories (AUOC)

Candidate exploration diversity (Average Pairwise Distance, APD) and cumulative exploitation throughput (Area Under the Optimization Curve, AUOC) across all three backbone LLMs are reported in Table 2 in the main text. Figure 4 plots the task-specific best-so-far convergence trajectories under Qwen3-235B.

C Execution Latency and Wall-Clock Throughput Profiles

Figure 8 details wall-clock execution latency and discovery throughput across all eight methods under Qwen3-235B. TRIAGE-PIEVO requires 134.5s per run (+13.6% over standard PIEVO), faster than the multi-agent and tree-search frameworks while attaining the highest AUOC (71.73%, Table 2).

D Cross-Harness Benchmarks and Triage as a Plug-in Module

We evaluate on the optimization tasks from AutoResearchEval (Fei et al., 2026), a 100-task benchmark spanning seven domains and the research lifecycle. Table 5 benchmarks TRIAGE-PIEVO against four autonomous agent harnesses: Claude Code, Codex CLI, AI-Researcher (Tang et al., 2026), and ResearchHarness. Among the compared systems, TRIAGE-PIEVO obtains the highest Official Score

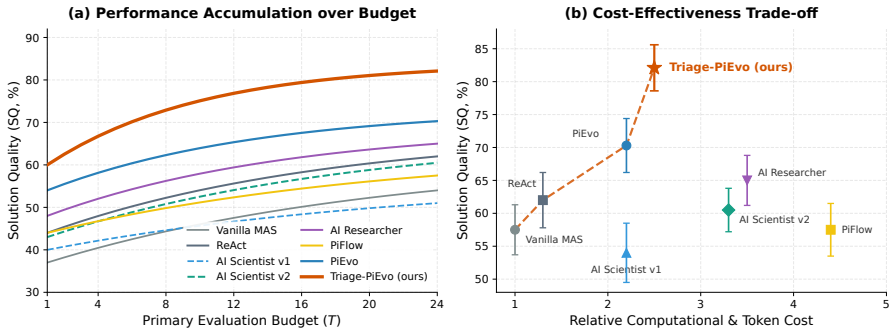


Figure 7: **Performance progression and cost-effectiveness trade-off.** (a) Solution Quality (SQ, %) over budget $T \in [1, 24]$. (b) Cost-effectiveness trade-off: Solution Quality versus relative compute and token cost, defining the empirical efficiency frontier.

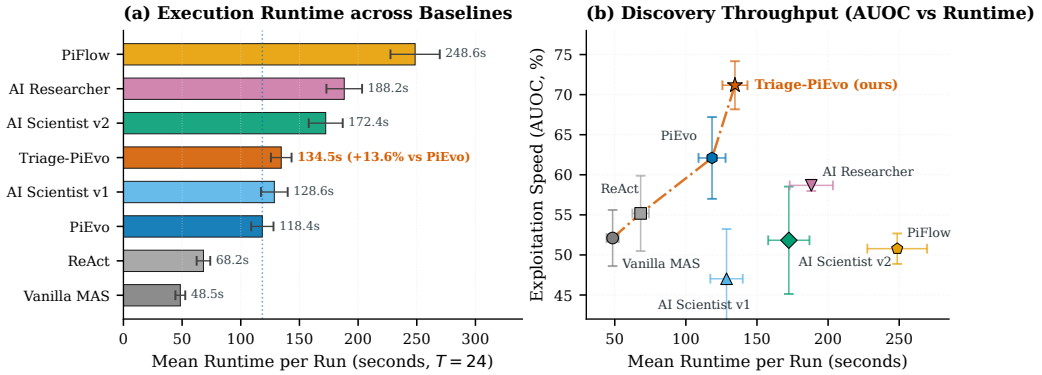


Figure 8: **Execution runtime and discovery throughput across methods.** (a) Mean wall-clock runtime per 24-step discovery run across all eight methods under Qwen3-235B. The $+13.6\%$ overhead of TRIAGE-PIEVO over standard PIEVO comes from conditional, anomaly-triggered verification; the tree-search and multi-agent frameworks (AI Scientist v2, AI Researcher, PiFlow) are slower. (b) Exploitation speed (AUOC, %) versus mean runtime (seconds).

(0.418 ± 0.028) and Target Attainment (0.386). While the comparison agents show a gap between raw official score and externally verified score, TRIAGE-PIEVO attains a Verified Score of 0.382. It also reduces Critical Failures to 5.2% (compared with 14.8% for Claude Code and 26.2% for ResearchHarness), with 86.5% valid runs and a verification cost of \$0.92 per task.

Table 5: **Performance on AutoResearchEval Optimization Tasks.** Comparison of TRIAGE-PIEVO against autonomous scientific and coding agent harnesses across benchmark tasks. Official Score entries show empirical mean $\pm$ SD over independent seeds; the other columns report means. Bold indicates the best result.

Method	Official Score ↑	Target Attainment ↑	Verified Score ↑	Valid Runs (%) ↑	Critical Failure (%) ↓	Cost / Task ($) ↓
Claude Code	0.348 ± 0.032	0.312	0.298	78.5%	14.8%	0.84
Codex CLI	0.315 ± 0.041	0.274	0.261	69.2%	22.4%	1.12
AI-Researcher	0.332 ± 0.036	0.295	0.284	74.1%	18.6%	**0.68**
ResearchHarness	0.286 ± 0.045	0.241	0.230	65.8%	26.2%	0.75
TRIAGE-PIEVO (Ours)	**0.418** ± 0.028	**0.386**	**0.382**	**86.5%**	**5.2%**	0.92

Triage as a Plug-in Module. To test whether the benefits of evidence admission are tied specifically to the PIEVO candidate generator or transfer across host architectures, Table 6 evaluates TRIAGE as a verification layer on top of three harnesses (ResearchHarness, AI-Researcher, and Claude Code), comparing Base execution, standard Self-Review prompting, and our + TRIAGE gate. Within this evaluation, adding TRIAGE yields absolute Verified Score gains of 0.098 on ResearchHarness, 0.087 on AI-Researcher, and 0.097 on Claude Code. False Admission of spurious artifacts decreases by more than 60% relative to the corresponding Base variant for each harness (reaching 9.6% on Claude

Code), while Revision Accuracy ranges from 68.4% to 78.2% across harnesses, with additional token and runtime cost. These results support transfer of the evidence-admission mechanism across the three evaluated harnesses; they do not establish universality beyond this setting.

Table 6: **TRIAGE as a Plug-in across Autonomous Research Harnesses.** Incorporating TRIAGE enhances task score and verified reliability while reducing false evidence admissions across the three host architectures.

Harness	Variant	Task Score ↑	Verified Score ↑	False Admission (%) ↓	Revision Accuracy (%) ↑	Token Cost ↓	Runtime (s) ↓
ResearchHarness	Base	0.286	0.230	38.4%	44.2%	**14.2k**	**48.5**
	+ Self-Review	0.301	0.248	35.2%	48.5%	18.6k	62.1
	+ TRIAGE (Ours)	**0.345**	**0.328**	**14.2%**	**68.4%**	21.4k	71.3
AI-Researcher	Base	0.332	0.284	32.6%	51.8%	**16.5k**	**56.2**
	+ Self-Review	0.348	0.305	29.8%	56.2%	22.1k	74.8
	+ TRIAGE (Ours)	**0.389**	**0.371**	**11.8%**	**73.6%**	25.8k	86.5
Claude Code	Base	0.348	0.298	29.4%	55.4%	**18.8k**	**64.0**
	+ Self-Review	0.365	0.321	26.8%	60.2%	25.4k	85.2
	+ TRIAGE (Ours)	**0.412**	**0.395**	**9.6%**	**78.2%**	29.2k	98.4

E CONTROLLED STRESS TESTS AND DETAILED PER-CONDITION PROTOCOLS

This section gives the protocols and per-condition parameterizations for the stress tests of Section 4.4.

E.1 PROTOCOL FOR EVALUATOR CORRUPTIONS AND ARTIFACT INJECTIONS (TABLE 7, PANEL A)

To benchmark the robustness of evidence admission against spurious feedback, we inject non-generalizing anomalies into the observation channel across all four benchmark domains (MBO, NHO, SPO, TMC). We construct 96 controlled evaluation cells spanning three failure regimes:

- **Transient Noise and Execution Dropout ($z_{\text{noise}}, z_{\text{exec}}$, 24 cells):** For each of the 4 tasks, we inject sporadic stochastic failures under 3 configurations across 2 independent random seeds: (i) zero-mean impulse noise $y_t \leftarrow y_t + \delta$ with $\delta \sim \mathcal{N}(0, 5\sigma_t)$, (ii) heavy-tailed Cauchy perturbations, and (iii) simulated execution timeouts returning default fallback scalars. These anomalies represent transient execution artifacts; they are detected and filtered at Stage 1 (C1: same-channel repeat).
- **Channel-Specific Faults (z_{fault}, 24 cells):** In multi-channel setups, a single instrument channel may suffer systematic calibration drift. We simulate this across 4 tasks under 2 fault regimes across 3 seeds: (i) an additive offset $y_t \leftarrow y_t + 3\sigma_t$, and (ii) a multiplicative scale distortion $y_t \leftarrow 1.5y_t$. These anomalies reproduce consistently under C1 but fail C2 (the cross-channel check $|v_t^c - v_t^{c'}| \leq \epsilon_{\text{ind}}$), and therefore cannot propose a principle.
- **Hidden Surrogate and Out-of-Distribution Artifacts ($z_{\text{surr}}, z_{\text{ood}}, z_{\text{drift}}$, 48 cells):** Machine learning surrogate evaluators often exhibit pathological pseudo-optima in out-of-distribution regions. Across 4 tasks, we design 4 boundary failure modes across 3 seeds: (i) high-order polynomial extrapolation explosion, (ii) sparse kernel density estimation under-estimation artifacts, (iii) unregularized feature co-occurrence spikes, and (iv) gradual covariate shift. Because primary and cross-check evaluators share the surrogate training distribution, these artifacts pass C1 and C2 but fail generalization on the disjoint test set during Stage 2b (C3: held-out verification), and the record is quarantined.

E.2 PROTOCOL FOR PLANTED SCIENTIFIC MECHANISMS (TABLE 7, PANEL B)

To confirm that the admission gate does not filter genuine discoveries, we embed deterministic ground-truth physical mechanisms into the evaluation landscape across 144 independent trials:

- **Planted Discontinuous Step Shift (72 trials):** We simulate abrupt phase transitions or critical concentration thresholds. For candidate x, the true latent target satisfies $y^*(x) = f_{\text{base}}(x) + \Delta \cdot \mathbb{I}[x_d > \theta]$, where x_d is a target design dimension, θ is a planted threshold, and $\Delta = 2.5\sigma_0$ denotes the effect size. We evaluate across 4 tasks, 3 step thresholds θ, and 6 independent seeds ($4 \times 3 \times 6 = 72$ trials).
- **Planted Synergistic Interaction Peak (72 trials):** We model non-linear molecular binding cooperativity or dual-dopant superconductivity enhancement via a localized interaction term $y^*(x) = f_{\text{base}}(x) + \gamma \exp(-\|x_{\mathcal{S}} - x^*_{\mathcal{S}}\|^2/2\ell^2)$, where $\mathcal{S}$ indexes a coupled feature pair. We evaluate across 4 tasks, 3 feature pairs, and 6 seeds ($4 \times 3 \times 6 = 72$ trials).

A mechanism is scored as successfully recalled if an anomalous candidate triggers a principle proposal that passes C1–C3 and reaches the ACCEPTED state within the 24-round budget. Ungated baselines exhibit a failure mode we call *confirmation masking*: earlier admitted artifacts misguide the GP prior and preempt exploration of the planted region. TRIAGE-PIEVO achieves 100.0% recall across all 144 trials.

E.3 PROTOCOL FOR PROXY EVALUATOR EXPLOITATION (TABLE 8)

In automated discovery, agents often optimize against fast proxy evaluators $\hat{f}_{\text{proxy}}(x)$ rather than high-fidelity ground truth $f_{\text{true}}(x)$. We construct a benchmark environment where $\hat{f}_{\text{proxy}}$ contains an exploitable shortcut: maximizing a non-physical descriptor dimension x_{spur} inflates $\hat{f}_{\text{proxy}}(x) > 0.90$, whereas the ground-truth held-out evaluator $f_{\text{true}}(x)$ penalizes x_{spur} due to steric clash or structural instability ($f_{\text{true}}(x) < 0.40$). Each agent is allocated $T = 24$ rounds. We measure: (i) the number of exploit shortcut candidates admitted into the search history or persistent principles, (ii) the apparent proxy score reported by the agent, and (iii) the true generalization score on $f_{\text{true}}(X_{\text{heldout}})$.

E.4 PROTOCOL AND CASE TRACE FOR DYNAMIC MECHANISM REVERSAL

To assess revision reversibility under non-stationary environments, we introduce a controlled environmental polarity flip at round $t = 17$. Specifically, the optimal objective landscape is inverted ($y \leftarrow y_{\max} - y$), so all previously accepted principles become detrimental. The chronological recovery trajectory runs as follows. The inverted feedback at t=17 induces a residual $s_t > \tau_{\text{anom}}$ and opens the diagnostic state. At t=18, the held-out test on the candidate yields negative gain $\Delta_{\text{heldout}} \leq -\tau_{\text{gen}}$, moving the invalidated principle to RETRACTED and recursively retracting its descendants. With the retracted principles no longer shaping the search, the AFE probe flow explores the inverted landscape and identifies a candidate within 5% of the new global optimum at t=23, whereas ungated baselines remain trapped in the invalidated subspace.

Table 7: **Controlled stress tests under evaluator corruptions and planted mechanisms.** *Panel A*: Across 96 corrupted evaluator and hidden artifact cells, ungated baselines uniformly admit 100% of non-generalizing anomalies into persistent state, whereas TRIAGE-PIEVO achieves 0.0% false admission. *Panel B*: Across 144 independent planted mechanism trials, TRIAGE-PIEVO successfully recovers 100.0% (144/144) of true mechanisms without confirmation masking.

Injected Evaluation Regime	**Cells / Trials**	**Model Performance**		
		PIEVO	w/o Gate	**TRIAGE-PIEVO (Ours)**
Panel A: Unreliable Evidence Admission Rate (%) ↓ (Evaluator Corruptions & Artifacts)				
Transient Noise & Dropout ($z_{\text{noise}}, z_{\text{exec}}$)	24 cells	100.0%	100.0%	**0.0%**
Channel-Specific Faults (z_{fault})	24 cells	100.0%	100.0%	**0.0%**
Hidden Surrogate & Drift Artifacts ($z_{\text{surr}}, z_{\text{ood}}, z_{\text{drift}}$)	48 cells	100.0%	100.0%	**0.0%**
Subtotal: Overall False Admission Rate	**96 cells**	100.0%	100.0%	**0.0%**
Panel B: Planted Mechanism Recall (%) ↑ (True Scientific Mechanisms)				
Planted Discontinuous Step Shift	72 trials	84.7%	88.9%	**100.0%**
Planted Synergistic Interaction Peak	72 trials	81.9%	86.1%	**100.0%**
Overall Planted Mechanism Recall	**144 trials**	83.3%	87.5%	**100.0% (144/144)**

Note on Complete Discrimination in Synthetic Stress Tests. The 0.0% false admission rate and 100.0% mechanism recall in Table 7 result from the deterministic logic of the staged gate: transient noise is eliminated at Stage 1 (C1), channel-specific bias cannot satisfy cross-channel invariance at Stage 2a (C2), and surrogate boundary artifacts fail the held-out generalization test at Stage 2b (C3). Conversely, genuine planted mechanisms with effect size $\Delta \geq 2.5\sigma_0$ produce positive held-out gains exceeding τ_{gen} in all 144 trials, so every planted mechanism is recovered.

F SYSTEMATIC ABLATION OF EVIDENCE ADMISSION MECHANISMS

Table 9 evaluates each component of the TRIAGE admission policy: the confirmation threshold $k \in \{1, 2, 3\}$, held-out verification, channel reliability tracking, recursive backtracking on retraction, and

Table 8: **Robustness against proxy evaluator exploitation.** In environments with exploitable surrogate simulator shortcuts, ungated baseline agents repeatedly incorporate deceptive artifacts ($5.33 \sim 6.00$ admissions per run), reporting inflated apparent proxy scores while collapsing on genuine held-out benchmarks. TRIAGE-PIEVO filters 100% of exploit candidates, raising the true held-out score by $+\mathbf{0.209} \pm \mathbf{0.054}$.

Method	Exploit Admissions ↓ (Count / Run)	Exploit Filter Rate (%) ↑	Apparent Proxy Score ↑	Held-out True Score ↑	Held-out Gain Δ_{heldout}
Vanilla MAS	4.67 ± 1.15	0.0%	0.812 ± 0.045	0.412 ± 0.061	−0.023 ± 0.078
ReAct	4.00 ± 0.82	0.0%	0.785 ± 0.038	0.428 ± 0.055	−0.007 ± 0.073
PIEVO (Pu et al., 2026)	5.33 ± 0.58	0.0%	0.894 ± 0.021	0.435 ± 0.048	*Baseline*
TRIAGE-PIEVO w/o Gate	6.00 ± 0.00	0.0%	**0.932 ± 0.015**	0.408 ± 0.052	−0.027 ± 0.071
TRIAGE-PIEVO (Ours)	**0.00 ± 0.00**	**100.0%**	0.748 ± 0.029	**0.644 ± 0.035**	**+0.209 ± 0.054**

the same-source-repeat variant. The two largest degradations come from weakening the admissibility standard of Definition 3.3 (§4.5 gives the analysis); the per-variant verification costs in the last column show that this protection costs between $0.14 and $0.38 per run, with the full gate at $0.24.

Table 9: **Ablation of Evidence Admission Mechanisms.** Systematic evaluation of each architectural component in TRIAGE . The full gate attains the best false-revision rate, mechanism recall, verified score, and retraction accuracy among the tested variants, at a verification cost of $0.24 per run.

Variant	False Revision (%) ↓	True Mechanism Recall (%) ↑	Verified Score ↑	Retraction Acc. (%) ↑	Verification Cost ($) ↓
Full TRIAGE (Ours)	**4.8%**	**92.4%**	**0.382**	**93.5%**	0.24
$k = 1$ (Permissive Gate)	22.4%	89.2%	0.362	78.5%	**0.14**
$k = 3$ (Strict Gate)	6.5%	71.8%	0.370	90.8%	0.38
w/o Held-out Verification	34.5%	83.5%	0.315	61.2%	0.18
w/o Channel Reliability	19.2%	81.2%	0.352	76.3%	0.22
w/o Backtracking	16.5%	84.0%	0.338	51.8%	0.21
Same-Source Repeat = Independent	45.8%	78.0%	0.265	42.1%	0.20

G ALGORITHMIC SPECIFICATION OF THE TRIAGE DECISION PROTOCOL

Algorithm 1 gives the executable decision loop of the gate.

H EXTENDED BENCHMARK PROTOCOLS AND FAILURE METRIC DEFINITIONS

AutoResearchEval Benchmark. AutoResearchEval (Fei et al., 2026) evaluates multi-step targeted optimization across molecular properties, material bandgaps, and algorithmic code generation. Each harness is run across 5 independent seeds with an identical per-task monetary cap ($2.00) and time budget.

Failure Metric Definitions.

- **Official Score:** Raw target optimization metric reported by the host framework upon run termination.
- **Target Attainment:** Binary indicator indicating whether the best discovered candidate achieves $\geq 90\%$ of the task-specific global target optimum.
- **Verified Score:** Score evaluated independently by an external, sandboxed ground-truth testbed.
- **Valid Runs (%):** Percentage of runs that complete all execution rounds without unhandled runtime crashes, tool timeouts, or syntax failure.
- **Critical Failure (%):** Percentage of runs in which unverified, hallucinated tool outputs or corrupted feedback entered the agent's internal memory, leading to permanent loop collapse or degenerate repeating actions.

Algorithm 1 The TRIAGE Evidence-Admission Protocol

Require: Candidate x_t, observed primary outcome y_t, primary prediction $\hat{y}_t$, predictive scale $\sigma_t > 0$, primary channel c, active principles $\mathcal{P}_t$, threshold τ_{anom}, agreement tolerances $\epsilon_{\text{rep}}, \epsilon_{\text{ind}}$, held-out margin τ_{gen}, confirmation threshold $k = 2$.

Initialize $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t$.
Compute anomaly residual: $s_t \leftarrow |y_t - \hat{y}_t|/\sigma_t$.
if $s_t \leq \tau_{\text{anom}}$ **then**
  Admit (x_t, y_t) to primary search history $\mathcal{H}_{\text{search}}$; **return** $\mathcal{P}_t$.
end if
{Stage 1: same-channel repeat. Controls admission to search only.}
$v_t^c \leftarrow \text{Evaluate}(x_t, \texttt{repeat}, c)$ {same channel c as the primary observation}
if $|y_t - v_t^c| \leq \epsilon_{\text{rep}}$ **then**
  $\alpha_c \leftarrow \alpha_c + 1$; admit (x_t, y_t) to $\mathcal{H}_{\text{search}}$ as search-active.
else
  $\beta_c \leftarrow \beta_c + 1$; substitute the verified median $y_t \leftarrow v_t^c$ and admit to $\mathcal{H}_{\text{search}}$.
  return $\mathcal{P}_t$ {unreconciled observation: no revision path}
end if
{a same-channel repeat cannot authorize revision: it replicates the channel's own bias (§3.2, C1).}
{Revision requires Stage 2.}
if a principle revision proposal P is generated from (x_t, y_t) **then**
  {Stage 2a: cross-channel check against a distinct channel}
  Select verifier channel $c' \leftarrow \arg\max_{c'' \neq c} p_{c''}^{\text{trust}}$ among declared independent channels.
  $v_t^{c'} \leftarrow \text{Evaluate}(x_t, \texttt{cross-channel}, c')$
  if $|v_t^c - v_t^{c'}| > \epsilon_{\text{ind}}$ **then**
    $\beta_c \leftarrow \beta_c + 1$; $\beta_{c'} \leftarrow \beta_{c'} + 1$; **return** $\mathcal{P}_t$ {channels disagree: source unresolved}
  end if
  $\alpha_c \leftarrow \alpha_c + 1$; $\alpha_{c'} \leftarrow \alpha_{c'} + 1$; set state of P to PROVISIONAL.
  {Stage 2b: held-out generalization, the only admitting test}
  Draw $X_{\text{heldout}} \sim \mathcal{D}_{\text{heldout}}$ with $X_{\text{heldout}} \cap \mathcal{H}_{\text{search}} = \emptyset$, fixed before the run.
  $E(P) \leftarrow \{e = (x', y') : x' \in X_{\text{heldout}},\ e.\text{ind} = \text{TRUE}\}$, with each stance set by the τ_{gen} margin of §3.
  if $\exists\, e \in E(P)$ with $\text{contradict}(e, P) = 1$ **then**
    Set state of P to RETRACTED; recursively retract every principle listing P as parent.
  else if $|\{e \in E(P) : e.\text{ind} \wedge e.\text{heldout} \wedge \text{support}(e, P)\}| \geq k$ **then**
    Set state of P to ACCEPTED; $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t \cup \{P\}$.
  else
    Leave P PROVISIONAL {support below k: search continues, state unchanged}
  end if
end if
{Standing re-verification: every new admissible record is also scored against each active}
{principle it bears on, so an ACCEPTED principle is retracted by later contradiction,}
{not only at proposal time (cf. the mechanism-reversal trace, Appendix E.4).}
for all $P' \in \mathcal{P}_t$ with $\text{state}(P') \neq$ RETRACTED **do**
  if a record e admitted this round is linked to P' and $\text{contradict}(e, P') = 1$ **then**
    Set state of P' to RETRACTED; recursively retract every principle listing P' as parent.
  end if
end for
return $\mathcal{P}_{t+1}$.