Discovery report for Critique Markets for Discovery: Turning Reviewer Feedback into a Resource Allocation Signal

Research Objective

Run this idea for me:
Critique Markets for Discovery: Turning Reviewer Feedback into a Resource Allocation Signal
TL;DR: Give your AI scientist a picky reviewer who pays "attention tokens" only for genuinely good ideas—then let the AI chase what the reviewer rewards. We'll embed an automated reviewer and treat its scores as a budget signal to schedule how the agent spends cycles on reading, coding, or synthesis, probing the observed linear scaling of findings in Kosmos.
Research Question: Can integrating automated critique and market-like allocation improve the marginal value of each additional Kosmos cycle, and where does the linear scaling of valuable findings begin to saturate?
Hypothesis: Turning reviewer feedback into a market signal that controls exploration-exploitation and cycle allocation will (i) increase novelty-adjusted accuracy per cycle and (ii) reveal a domain-dependent scaling law and saturation point earlier than naïve linear extrapolation suggests.
Experiment Plan: - Setup:
Base system: Kosmos-like multi-agent loop. Automated reviewer: AI Scientist's reviewer calibrated on held-out human reviews and rubric-anchored scores (novelty, rigor, clarity, evidence). Critique market: each hypothesis/analysis "bids" for cycles; reviewer returns "payouts" that drive a bandit-style scheduler to allocate the next cycles to the most promising lines of inquiry. Data/Materials: Two domains with different evidence dynamics: (1) plant science (Aleks case; structured data, interpretable models), (2) materials (text-heavy with code experiments). Optionally include a qualitative synthesis task (ethics literature) to test generality. Measurements: Slope of valuable findings vs. cycles with and without critique markets. Reviewer-human agreement; stability of scores over time. Saturation detection: cycles at which marginal payouts drop below threshold; compare across domains. Expected Outcomes: Improved novelty-adjusted accuracy per unit time. Empirical characterization of scaling (linear in early regime; sublinear plateau) and identification of intervention points (e.g., literature refresh or role rotation) that delay saturation. References: ['Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv.org.', 'Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., et al. (2025). Kosmos: An AI Scientist for Autonomous Discovery. Preprint.', 'Jin, D., Gunner, N., Carvajal Janke, N., Baruah, S., Gold, K., & Jiang, Y. (2025). Aleks: AI powered Multi Agent System for Autonomous Scientific Discovery via Data-Driven Approaches in Plant Science. arXiv.org.', 'Eger, S., Cao, Y., D'Souza, J., Geiger, A., Greisinger, C., Gross, S., Hou, Y., Krenn, B., Lauscher, A., Li, Y., Lin, C., Moosavi, N., Zhao, W., & Miller, T. (2025). Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. arXiv.org.', 'Reddy, C. K., & Shojaee, P. (2024). Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges. AAAI Conference on Artificial Intelligence.']

Summary of Discoveries

Discovery 1: Calibrated Automated Review as a Scalar Payout for Discovery Markets

Calibrated automated review can be converted into a single, quantitative "payout" that governs how an AI scientist allocates its attention across hypotheses, reading, coding, and synthesis. By fusing ensemble, retrieval-grounded, pairwise reviewers with rubric-anchored scoring, non-compensatory rigor gating, and domain-specific rigor metrics for code and simulation, the payout becomes stable enough to drive a market-like bandit scheduler and to probe scaling and saturation of valuable findings.

Discovery 2: Critique-Market Scheduling: Bandit-Based Allocation of Scientific Cycles

Critique-Market Scheduling converts automated reviewer feedback into an explicit payout signal that a bandit scheduler uses to allocate scarce “attention tokens” across hypotheses and tasks. By coupling rubric-anchored ensemble reviewers with market-like bidding and bandit allocation, the approach aims to raise novelty-adjusted accuracy per cycle and to empirically map where linear gains in valuable findings saturate across domains.

Discovery 3: Online Saturation Detection and Memory Control via Bayesian Change Points

This work proposes to turn an automated reviewer’s payouts into a market signal that both detects when the return on additional cycles begins to saturate and adaptively controls the agent’s memory and exploration. The core mechanism is online Bayesian change point detection over the payout time series, with posterior-triggered resets or discounts that steer a bandit-style scheduler toward promising lines of inquiry and away from plateaued ones.

Discovery 4: Meta-Controlled Interventions to Sustain Discovery Yield

This work proposes a critique market that converts automated reviewer scores into a market-like “payout” signal to steer an AI scientist’s compute across reading, coding, and synthesis. A meta-controller monitors the payout time series to detect saturation and triggers principled interventions—spanning exploration boosts, representation refresh, and posterior discounting—to restore marginal value per cycle.

Calibrated Automated Review as a Scalar Payout for Discovery Markets

Summary

Calibrated automated review can be converted into a single, quantitative "payout" that governs how an AI scientist allocates its attention across hypotheses, reading, coding, and synthesis. By fusing ensemble, retrieval-grounded, pairwise reviewers with rubric-anchored scoring, non-compensatory rigor gating, and domain-specific rigor metrics for code and simulation, the payout becomes stable enough to drive a market-like bandit scheduler and to probe scaling and saturation of valuable findings.

Background

Closed-loop, multi-agent research systems increasingly automate literature search, hypothesis generation, and analysis, but they rarely transform reviewer critique into explicit signals that allocate scarce compute and researcher cycles. Meanwhile, meta-science indicates that iterative review markedly improves downstream quality and throughput, and automated reviewers can approximate human judgments in selective contexts. To test whether critique can do double duty—as evaluation and as a market-like budget allocator—this work builds a calibrated scalar payout from automated reviews and uses it to control exploration–exploitation in Kosmos-style loops, with the goal of increasing novelty-adjusted accuracy per cycle and mapping where linear gains saturate.

Results & Discussion

The central discovery is that automated critique can be made both reliable and actionable by (i) upgrading review to ensemble, retrieval-grounded, pairwise comparison with input-order randomization and (ii) translating rubric-anchored aspect scores into a single scalar payout for scheduling. Evidence from multi-model reviewer pipelines already shows large quality gains between initial and revised rounds and formalizes gatekeeping via majority voting (e.g., 3/5 accepts) in aiXiv-style workflows, highlighting that reviewer–author feedback is a high-leverage control signal [r0, zhang2025a]. Automated reviewers have been used inside closed loops as reward or selection models: The AI Scientist validates an Automated Paper Reviewer and argues that scaling search-and-filter yields better papers at low cost; CycleReviewer reduces mean absolute error versus individual human score predictions by 26.89%, and AI-Researcher reports perfect alignment with ICLR final decisions on a paired-review experiment, strengthening the case for rubric-anchored, pairwise scoring as a control signal [r0, tang2505]. Complementarily, ensemble architectures with pairwise comparisons and input-order randomization improve stability over single-model direct scoring, mitigating verbosity and order biases and increasing agreement with humans [r3, gao2024]. These gains coexist with known risks—hallucination, prompt sensitivity, calibration drift, and fragile novelty detection in natural sciences—so retrieval grounding, ensembling, and debiasing are treated as first-class components of the payout pipeline [r0, mann2025].

Constructing the scalar payout starts from a unified, tri-axial rubric spanning novelty, rigor, and clarity, harmonized from published reviewer checklists and public peer-review corpora; the literature supports these dimensions but emphasizes heterogeneity and the need for validation of a consolidated rubric across fields [r11, zhang2025]. Novelty is stabilized with a primary embedding-distance signal (e.g., cosine distance and percentile scores) that correlates with self-reported novelty and future impact, complemented by knowledge-graph topology (new links among concepts) and surprisal-based measures to capture orthogonal facets of originality [r5, zhao2025]. To make the payout robust and interpretable, standard composite-indicator practice is applied: direction alignment, data-type-aware normalization (Boolean constraints as 0/1 or vetoes; ordinal/categorical mapped to discrete classes and rescaled; continuous min–max or percentile scaling), and aggregation with transparent weighting [r24, brito2018]. Crucially, rigor acts as a non-compensatory gate using established functional forms from decision theory: an indicator-based threshold I[rigor T] that masks downstream contributions when rigor falls below T, or an outranking-style veto to enforce hard constraints; within-gate, weighted linear or geometric means control compensability [r22, r24, garcia2019]. Sensitivity analysis over weights, thresholds, and normalizers closes the loop to quantify robustness of the payout under plausible perturbations [r24, brito2018].

Domain-tailored rigor metrics make the payout actionable across heterogeneous workflows. For Python data-analysis scripts and notebooks typical of plant science and text-heavy code experiments, the rigor composite draws from three automatable classes: (1) static analysis/style linting (e.g., Pylint, Flake8, Hyperstyle; notebook-aware tools such as Julynter and Matroskin) to flag unused variables, formatting violations, and potential bugs; (2) execution-environment verification (e.g., ReproZip, explicit version pinning with environment managers) to ensure dependency completeness and consistency; and (3) output stability checks via test assertions and notebook-execution frameworks to verify consistent results across runs [r19, wang2020a]. Integration must account for notebook hidden state and non-linear cell execution, so notebook-aware static and dynamic checks are included and can be treated as Boolean vetoes or penalized factors in the rigor subgroup [r19, grotov2022]. For high-throughput materials simulations, established workflow frameworks (AiiDA/ACWF, atomate/atomate2) already implement input-schema and metadata validation, protocolized physical-plausibility checks (e.g., k-point meshes, smearing, symmetry), automated convergence and error handling (e.g., custodian), and parsers that surface termination and threshold criteria; these yield a machine-computable rigor score, optionally augmented by physically grounded metrics (, ϵ, ν) for equations of state and provenance completeness [r54, schultz2011]. Parameter thresholds are property- and engine-specific and must be protocolized per target, with uncertainty propagation where fits are involved to avoid overconfidence [r54, bosoni2024]. Together, these domain-specific rigor composites plug into the global rubric as hard constraints (veto) and as low-compensation subaggregates, ensuring that novelty cannot wash out methodological fragility [r22, r24, garcia2019].

With a calibrated payout in hand, a bandit-style scheduler spends “attention tokens” on hypotheses and analyses that deliver the highest marginal returns, favoring bandits over sample-hungry RL to keep training data and cost requirements low [r0, zhu2025]. The market allocates cycles across reading, coding, and synthesis, and updates estimates using reviewer “payouts” after each tranche; retrieval-grounded, ensemble, pairwise reviewers stabilize these payouts, while hybrid RAG+graph retrieval specifically targets novelty fidelity under domain drift and adversarial prompts [r0, zhang2025]. Evaluation focuses on (i) the slope of valuable findings versus cycles with and without critique markets, (ii) reviewer–human agreement and stability over time (e.g., Cohen’s and Spearman correlation), and (iii) saturation detection when marginal payouts fall below a threshold, comparing plant science (structured tabular pipelines with predictive metrics such as R^2, RMSE, F1) and materials simulation domains [r0, r11, r13, jin2025]. The literature anticipates early linear gains followed by a sublinear plateau and identifies intervention points—literature refresh, role rotation, or rubric retuning—that can delay saturation; head-to-head Kosmos/Aleks evaluations under explicit critique markets remain a key gap to close [r0, zhang2025a]. Finally, because peer-review scores can poorly predict long-term impact and multi-agent reviews can exhibit social biases, the system incorporates human green-light checkpoints for high-risk allocations and runs weight/threshold sensitivity analyses to ensure that performance does not hinge on brittle rubric settings [r0, zhang2025].

Trajectory Sources

Trajectory r0: Overview and motivation. Turning automated critique into an explicit allocation signal aims to fix two linked bottlenecks: scarce, noisy human review and uncalibrated exploration in autonomous discovery loops. Multi-agent, closed-loop research systems now exist, but they rarely convert reviewer judg...

Trajectory r3: Ensemble architectures that incorporate pairwise comparisons with input-order randomization produce more stable and reliable review scores than single-model direct scoring approaches (li2024 pages 20-22).

Trajectory r5: Computational measures based on semantic distance—particularly those employing word embeddings and cosine similarity—effectively quantify the novelty of scientific outputs, while complementary approaches using knowledge graph topology and surprisal analysis show promising potential (shibayama2021mea...

Trajectory r11: Partially supported: the meta-science literature and several public peer-review corpora encode novelty/rigor/clarity as explicit aspects, but these signals are heterogeneous, often heuristic, and lack a single validated rubric; a unified operational rubric is feasible only with additional harmonizat...

Trajectory r13: The Aleks system and related studies confirm that plant science discovery workflows rely on structured, predominantly tabular multi-omics and phenomics datasets analyzed with Python-based statistical and machine learning libraries to yield predictive models and statistical associations. (jin2025alek...

Trajectory r19: The literature confirms that established open-source tools—combining static code analysis, execution environment verification, and output stability checks—can be integrated to automatically generate rigor metrics for Python-based scientific scripts using pandas and scikit-learn (samuel2024computatio...

Trajectory r22: The literature confirms that explicit functional forms—such as indicator functions implemented in CART and threshold parameters used in outranking methods—can be adapted to allow a "base" metric to gate the contribution of a "performance" metric. (krupnick2006 pages 44-47, badea2011composite...

Trajectory r24: Supported: mainstream MCDA/composite-indicator literature specifies stepwise protocols that normalize Boolean, categorical, and continuous inputs to a common scale and aggregate them (typically via weighted linear or geometric means) into a single scalar suitable for reward design (moreira2021supple...

Trajectory r54: The hypothesis is supported: established HT frameworks provide automatable input-schema validation, physical-plausibility checks, and output-parsing/convergence verification, enabling a composite "Rigor" score constructed from these metrics. (bosoni2024 pages 3-4, schultz2011nuclearenergy...

Critique-Market Scheduling: Bandit-Based Allocation of Scientific Cycles

Summary

Critique-Market Scheduling converts automated reviewer feedback into an explicit payout signal that a bandit scheduler uses to allocate scarce "attention tokens" across hypotheses and tasks. By coupling rubric-anchored ensemble reviewers with market-like bidding and bandit allocation, the approach aims to raise novelty-adjusted accuracy per cycle and to empirically map where linear gains in valuable findings saturate across domains.

Background

Autonomous discovery systems now generate and refine hypotheses, read literature, write code, and synthesize results, but they still lack principled, data-driven mechanisms for deciding where to spend the next unit of compute or researcher attention. Human peer review is a strong control signal yet is too scarce, slow, and noisy to drive fine-grained scheduling, while automated reviewers are increasingly robust but remain underused as explicit allocation signals. A critique market operationalizes reviewer judgments as a market price, closing the loop between critique and computation so that exploration–exploitation decisions are driven by measured scientific value rather than static heuristics or naïve round-robin scheduling.

Results & Discussion

The discovery formalizes a bid–allocate–payout loop that turns automated critique into an allocation signal and embeds it inside a Kosmos-like multi-agent research cycle. The gap is clear: closed-loop, multi-agent systems and automated reviewers already improve quality and throughput at low marginal cost, yet almost none translate reviewer judgments into principled resource scheduling or quantify scaling laws of "valuable findings vs. cycles" across domains [r0, lu2024]. Prior work shows that ensemble LLM review, retrieval grounding, and iterative revision significantly upgrade proposals and manuscripts, with examples including large acceptance jumps after revision in aiXiv-style pipelines and near-human or even perfect alignment signals in selective settings (e.g., CycleResearcher reduces MAE versus individual human score predictions by 26.89% and AI-Researcher reports perfect alignment with ICLR final decisions on a paired-review experiment), motivating the use of reviewer scores as a learnable control signal for allocation [r0, weng2024, tang2505]. At the same time, novelty calibration drifts, hallucination risks, and social biases can destabilize naïve use of review scores, underscoring the need for robust scoring and continual calibration if those scores are to serve as payouts [r0].

The payout is a single scalar computed from a multi-dimensional rubric (e.g., novelty, rigor, clarity, evidence) using linear scalarization with domain-tuned weights, a pragmatic baseline from multi-objective RL that enables direct use by bandit algorithms while supporting outer-loop weight selection or adaptive elicitation when preferences shift [r2, hayes2021]. To improve score reliability, the market's reviewer is an ensemble that uses pairwise comparisons with input-order randomization and chain-of-thought style prompts, which consistently increases alignment with human judgments and reduces variance compared with pointwise single-model scores [r3, li2024, wang2025canllmsreplace, gao2024]. This produces a calibrated "score-per-cycle" objective for each task that feeds the scheduler. Because peer-review scores correlate imperfectly with long-term impact, the scalar payout is explicitly framed as a short-horizon control signal for allocation rather than a proxy for downstream citations or societal value, and it is monitored for calibration drift via reviewer–human agreement and stability audits over time [r0].

Operationally, task agents submit bids that encode requested cycles, SLOs, predicted runtime/latency risk, marginal cost, and expected reviewer-score-per-cycle; the scheduler applies feasibility filtering and then allocates cycles using a bandit-preprocessed choice rule (e.g., epsilon-greedy over a TOPSIS-like ranker or direct UCB on score-per-cycle), issues short leases to prevent double-booking, and routes realized reviewer scores back as payouts for online updates and retrospective scoring of bid predictions [r27, thiyagaraj2021, bhatt2025, ehsanfar2020]. Arms are standardized across domains as discrete research operations—hypothesis generation; literature/data retrieval; experiment/simulation design; code/simulation execution; and result synthesis—so that the same scheduler spans plant science data analysis and materials simulations while permitting domain-specific tuning [r23, zhou2025, coley2020, hao2025]. Because reward landscapes in research are nonstationary and many tasks return delayed scores, the scheduler employs advanced bandits (e.g., modified Thompson Sampling, contextual UCB) and delay-aware reductions (ABBD, SBBD) with regret guarantees without surrogate imputation, ensuring that exploration–exploitation adapts to drift and asynchronous completions [r9, r14, qin2202, liu2019, joulani2012].

The learning regime includes a pure-exploration burn-in followed by adaptive stopping rules to transition into exploitation when confidence bounds or information-accumulation criteria indicate sufficient model identifiability; this reduces early over-commitment to spurious reviewers or arms and is grounded in analyses from logistic and non-linear bandits [r41, jun2021improvedconfidencebounds, rajaraman2024]. To close the loop under delayed, terminal payouts (e.g., a synthesis note scored only at the end), credit assignment redistributes the reviewer reward across the action sequence using principled heuristics (last-touch, uniform split, Shapley-like removal effect) or model-based counterfactual decompositions, enabling the bandit to update value estimates for upstream arms such as literature retrieval or experiment design [r103, li2025a, ferret2022]. Evaluation targets two domains with different evidence dynamics—plant science (Aleks-style, structured/interpretables) and materials (text-heavy with code experiments)—and quantifies (i) slopes of valuable findings vs. cycles with and without the critique market, (ii) reviewer–human agreement and score stability, and (iii) saturation points where marginal payouts fall below a domain-specific threshold; where saturation emerges, interventions such as literature refresh or role rotation are triggered and audited for their ability to restore linear gains [r0, zhang2025a]. Collectively, the critique market reframes automated review from a gatekeeper into a continuous pricing signal for attention, providing a testable path to increase novelty-adjusted accuracy per unit time and to map early linear and late sublinear regimes of discovery yield across autonomous science workflows [r0, r23].

Trajectory Sources

Trajectory r0: Overview and motivation. Turning automated critique into an explicit allocation signal aims to fix two linked bottlenecks: scarce, noisy human review and uncalibrated exploration in autonomous discovery loops. Multi-agent, closed-loop research systems now exist, but they rarely convert reviewer judg...

Trajectory r2: A single scalar payout value can be constructed from multi-dimensional reviewer scores using a weighted function, provided that the optimal weights are elicited or learned in a domain-specific manner (hayes2021 pages 12-14, hayes2021 pages 14-15).

Trajectory r3: Ensemble architectures that incorporate pairwise comparisons with input-order randomization produce more stable and reliable review scores than single-model direct scoring approaches (li2024 pages 20-22).

Trajectory r9: Advanced MAB algorithms—specifically modified Thompson Sampling and contextual UCB variants designed to handle delayed feedback and dynamic reward structures—are better suited for scientific discovery than simple MAB approaches (qin2202 pages 1-2, qin2202adaptivityandconfound...

Trajectory r14: Our survey confirms that the literature indeed offers concrete MAB algorithms—such as ABBD and SBBD—that handle unknown and variable reward delays without resorting to surrogate imputation or discounting, making them directly applicable for scheduling computational tasks with uncertain completion ti...

Trajectory r23: The reviewed literature confirms that discrete action spaces—spanning hypothesis generation, literature retrieval, experimental and simulation design, code/simulation execution, and result synthesis—are well documented, supporting our hypothesis that these operational categories can form candidate a...

Trajectory r27: Yes—the surveyed agent-based scheduling and auction/bandit literature provides explicit, stepwise protocols for bidding, allocation, and reward routing that can be directly adapted to a “critique market.”

Trajectory r41: Our findings support the hypothesis that the optimal duration of a pure-exploration burn-in phase in multi-armed bandits can be determined either a priori from known problem parameters or adaptively using real-time statistical criteria (jun2021improvedconfidencebounds pages 4-6, hanUnknownyearimprov...

Trajectory r103: The literature confirms that both reinforcement learning and multi-armed bandit research offer formal credit-assignment strategies—including heuristic, model-based, and specialized bandit methods—that update value estimates for multi-stage action sequences leading to a single delayed terminal reward...

Online Saturation Detection and Memory Control via Bayesian Change Points

Summary

This work proposes to turn an automated reviewer's payouts into a market signal that both detects when the return on additional cycles begins to saturate and adaptively controls the agent's memory and exploration. The core mechanism is online Bayesian change point detection over the payout time series, with posterior-triggered resets or discounts that steer a bandit-style scheduler toward promising lines of inquiry and away from plateaued ones.

Background

Autonomous "AI scientist" systems now coordinate reading, coding, and synthesis across many cycles, but they lack principled mechanisms to recognize when a research thread's value is flattening and to reallocate effort accordingly. Embedding an automated reviewer provides a dense reward signal tied to novelty, rigor, and evidence, enabling critique markets in which hypotheses bid for compute and are funded according to performance. Detecting the onset of diminishing returns and modulating memory in real time are therefore key to improving the marginal value of additional cycles and to characterizing when linear scaling of valuable findings gives way to sublinear plateau.

Results & Discussion

This discovery formalizes saturation within critique markets as a sequential change in the reviewer payout process and shows how to trigger adaptive scheduling when the slope of cumulative valuable findings begins to flatten. The literature supports two complementary strategies to locate this transition: (i) online change point detection (CPD) that flags a shift from growth to plateau even under nonstationary and heteroscedastic conditions (e.g., dynamic-programming frameworks such as cp3o and G2CD, as well as Bayesian and penalized-spline approaches), and (ii) fitting sublinear learning-curve models (logistic, power-law) to anticipate asymptotic behavior; both families are evaluated via simulation RMSE and change-point scoring and are suitable for streaming deployment on the payout time series [r6, wenyu2020, aminikhanghahi2017, mohr2024, vianna2024, viering2022]. Because noise and long-memory dynamics can confound either method alone, the discovery recommends a hybrid: use online CPD for real-time alarms and maintain a concurrent sublinear fit as a secondary check to validate plateau onset in critique markets [r6, aminikhanghahi2017].

The trigger itself is grounded in Bayesian online CPD: maintain a run-length posterior $\mathrm{P}(\mathrm{r}_\mathrm{t}|\mathrm{z}_1\mathrm{:t})$ over payout regimes and fire when the posterior mass on a new regime $\mathrm{P}(\mathrm{r}_\mathrm{t} = 0|\cdot)$ exceeds a threshold, or when argmax run-length shifts, yielding immediate, posterior-based decisions suitable for meta-control of scheduling [r51, lin2023]. Restarted BOCPD variants provide non-asymptotic guarantees on false alarms and detection delay, enabling Average Run Length (ARL)–constrained thresholding for plateau detection; streaming/constant-memory versions support high-throughput operation, and deployments in network, asset, and clinical model surveillance demonstrate practicality of posterior-to-trigger mappings in real systems [r51, alami2020, wang2022a, murph2023]. Bandit-augmented quickest detection further shows how sampling policies can preserve ARL while reducing time-to-alarm, an important consideration when payouts are costly to obtain and the scheduler must balance exploration and confirmation [r51, zhang2023b].

To convert alarms into better allocation, the discovery links plateau detection to memory control in the scheduler via Thompson sampling (TS) mechanisms that have explicit, validated update rules. Detector-driven resets set posteriors back to diffuse priors (e.g., $\alpha{=}\beta{=}1$ for Bernoulli/Beta), maximally inflating variance and forcing exploration; partial refresh applies multiplicative decay to sufficient statistics ($\mathrm{S}_\mathrm{i} \leftarrow \gamma\mathrm{S}_\mathrm{i}$, $\mathrm{N}_\mathrm{i} \leftarrow \gamma\mathrm{N}_\mathrm{i}$), trading graceful forgetting for responsiveness; continuous discounting (Discounted TS) updates all arms each round with a fixed γ to manage effective concentration; and sliding-window TS limits its posteriors to the last τ observations to handle drift, with all variants supported by concrete pseudocode [r60, ghatak2022, ghatak2021, zaid2020, trovo2020, raj2017]. Beyond discrete alarms, a second contribution is to treat the Bayesian run-length posterior as a continuous control signal for discount, drawing on OU-style frameworks that map change probabilities to smoothly varying forgetting rates; while such posterior-to-γ mappings are conceptually well-motivated, most validation is simulation-based and context-dependent [r64, duranmartin2024]. The survey record also highlights an important gap: outside the constant-hazard equivalence $\gamma = 1 - \mathrm{h}$ (EWMA), explicit formulas γ_t = f(P(change at t)) are not standard, making critique markets a novel testbed for evaluating γ_t = $1 - \mathrm{P}(\mathrm{r_t}{=}0|\cdot)$ and γ_t functions of $\mathrm{E}[\mathrm{r_t}|\cdot]$ against detector-triggered resets and fixed discount baselines [r73, pasturel2020].

Finally, the system design adopts a hybrid nonstationary bandit to schedule cycles: passive memory fading (discount or windows) handles gradual drift in reviewer payouts, while active CPD fires when abrupt plateau or shock is detected; this combination is repeatedly shown to outperform either mechanism alone across environments with mixed nonstationarity, as measured by cumulative reward and dynamic regret [r92, cavenaghi2021, komiyama2107]. In practice, the scheduler treats hypotheses or analyses as arms whose “bids” are inferred from recent payouts and uncertainty, allocates cycles to maximize expected novelty-adjusted accuracy per unit time, and uses CPD alarms to trigger role rotation, literature refresh, or memory discounting when marginal payout falls below a threshold; sublinear curve fits operate in the background to corroborate and localize saturation [r6, wenyu2020, mohr2024]. Thresholds are selected under ARL constraints to stabilize false alarms, while the slope of valuable findings versus cycles is tracked with and without the critique market to quantify improvement and to expose domain-dependent early linear regimes and sublinear plateaus [r51, alami2020]. Together, posterior-triggered alarms, validated discount/reset controls, and hybrid scheduling convert reviewer feedback into a principled resource-allocation signal that should increase novelty-adjusted accuracy per cycle and reveal earlier, domain-specific saturation points than naïve linear extrapolation would suggest [r6, r51, r60, r92].

Trajectory Sources

Trajectory r6: The literature confirms that the saturation point can be effectively identified either by applying advanced change point detection algorithms or by fitting sublinear growth models such as logistic or power-law curves to the cumulative data (wenyu2020 pages 10-13, aminikhanghahi2017as...

Trajectory r51: The literature supports the hypothesis: online Bayesian changepoint detectors can drive real-time triggers by monitoring the posterior probability of a new changepoint and firing when that probability crosses a threshold. (lin2023 pages 2-3, alami2020restartedbayesianonlin...

Trajectory r60: Yes—the literature provides explicit algorithms and pseudocode that modulate Thompson-Sampling posteriors to control exploration by resetting, discounting, or windowing sufficient statistics, often triggered by external change-detection signals. (ghatak2022...

Trajectory r64: The evidence supports that Bayesian run-length posteriors can be used as inputs to continuous mapping functions that modulate discount factors, enabling graceful forgetting in adaptive filtering and bandit algorithms (duranmartin2024 pages 9-11, duranmartin2024 pages 11-13)...

Trajectory r73: The hypothesis is only partially supported: the literature provides an explicit linear mapping $\gamma = 1 - \mathrm{h}$ for constant hazard, but we found no explicit simple functions that map posterior change-point signals (e.g., BOCD run-length or P(change at t)) to a time-varying discount $\gamma(\mathrm{t})$ or window size $\tau(\mathrm{t}$...

Trajectory r92: Empirical studies show that hybrid non-stationary MAB algorithms—those combining passive memory-fading mechanisms with active change detection—tend to outperform methods relying exclusively on either sliding-window/discounting or change-point detection when facing both gradual drift and discrete cha...

Meta-Controlled Interventions to Sustain Discovery Yield

Summary

This work proposes a critique market that converts automated reviewer scores into a market-like “payout” signal to steer an AI scientist’s compute across reading, coding, and synthesis. A meta-controller monitors the payout time series to detect saturation and triggers principled interventions—spanning exploration boosts, representation refresh, and posterior discounting—to restore marginal value per cycle.

Background

Autonomous discovery systems are increasingly organized as multi-agent loops that propose, test, and synthesize hypotheses while tracking a stream of intermediate evidence and payoffs. Early progress often scales roughly linearly with compute before stalling as search dynamics, representations, or memory induce plateaus. The central challenge is to detect the onset of diminishing returns in real time and to intervene in ways that reconfigure exploration–exploitation, representation, and memory so that discovery yield per unit time is sustained without wasting cycles.

Results & Discussion

The core idea is to treat reviewer feedback as a scarce resource signal: each hypothesis or analysis “bids” for cycles, an automated reviewer returns rubric-anchored scores (e.g., novelty, rigor, clarity, evidence), and their composite becomes a payout that a bandit scheduler maximizes under a fixed cycle budget. The payout time series serves two roles: it drives short-horizon allocation via a bandit policy (e.g., Thompson Sampling) and supplies the meta-level state for longer-horizon control of interventions that reshape the search dynamics [r60, ghatak2022, zaid2020]. To make saturation operational, the system maintains both (i) a cumulative payout curve and (ii) rolling slopes of “valuable findings vs. cycles,” and declares diminishing returns when change-point detectors identify a shift from growth to plateau or when sublinear growth models (e.g., logistic, power-law) fit the cumulative curve with superior predictive score relative to linear baselines, approaches shown to work under nonstationary and heteroscedastic noise [r6, wenyu2020, aminikhanghahi2017, mohr2024, viering2022]. Known challenges—noise, long-memory, and parameter identifiability—are addressed by hybridization: ensemble change-point detectors validate sublinear fits before triggering a saturation flag, which reduces false alarms in difficult time series [r6, aminikhanghahi2017].

Once a saturation flag is raised, the decision to intervene is framed as a meta-level Markov decision process whose state concatenates engineered payout features (moving averages, derivatives, volatility) with a recurrent belief that summarizes history, enabling approximate Markovian credit assignment across long horizons [r34, cho2412, lek2023]. The meta-action space is a discrete set of temporally extended interventions, augmented with an explicit “do nothing” option; the meta-reward is the area-under-payout improvement over a fixed horizon following an intervention, optionally regularized by an intervention cost to avoid over-triggering [r34, fan2023]. A hierarchical controller operationalizes this setup: a slow-cadence DDQN or policy-gradient meta-controller selects interventions, while the low-level policy conditions on the chosen macro-action to execute primitives; this separation is standard in hierarchical RL and supports stable training and long-horizon credit assignment [r34, cho2412, sun2019]. For selecting among a small, discrete intervention set with low-dimensional time-series states, hybrid bandit–Bayesian optimization methods such as BOHB provide strong anytime performance by combining successive halving with density-estimation guidance, including kernels tailored to categorical choices, making them well suited to meta-level “best intervention” selection under budget constraints [r69, falkner2018].

The intervention library instantiates automatable strategies with demonstrated efficacy for escaping local optima and sustaining progress. Novelty- or quality-diversity–driven exploration increases the weight on behavioral novelty when progress stalls, maintaining an archive and adaptively reweighting reward–novelty to avoid deception; meta-populations with periodic extinction or resets reallocate compute toward promising lineages while mitigating initialization bias; non-elitist acceptance with temperature control allows temporary regressions and tunes exploration near the error threshold; iterated local search schedules controlled perturbations or heavy-tailed jumps followed by local improvement; and periodic representation refresh rethinks objectives or encoders to escape policy lock-in [r28, conti2018, coiffard2025, oliveto2018, dang2021, ochoa2019, cuccu2011]. Complementing these, a re-annealing schedule provides a sharp, temporary increase in exploration "temperature" followed by gradual decay, a well-studied mechanism in simulated annealing that can be mapped to widening posterior variance or modifying pseudo-counts in the bandit, thereby forcing escape from local optima when stagnation is detected [r74, ingber1989, battiti2009]. Two principled, low-level modulation knobs let the meta-controller tune exploration–adaptation continuously: (i) power discounting in Gaussian dynamic linear models inflates the prior covariance via $\mathrm{R_t} = \mathrm{G_t}$ C_{t−1} $\mathrm{G_t} /\delta$ with $\mathrm{W_t} = \mathrm{G_t}$ C_{t−1} $\mathrm{G_t}$ (δ^{-1} − 1), so smaller δ accelerates adaptation; multivariate extensions separate coefficient and covariance discount factors for finer control [r70, nobre2001, fisher2020], and (ii) Thompson Sampling variants that reset or multiplicatively discount sufficient statistics on change detection (e.g., KS-triggered resets, per-arm discounting, or sliding-window posteriors) explicitly modulate posterior concentration to boost exploration only where needed [r60, ghatak2022, ghatak2021, zaid2020, trovo2020].

Evaluation focuses on whether critique markets improve novelty-adjusted accuracy per cycle and on empirically characterizing where linear scaling yields to sublinear plateaus. The primary metrics are: (1) slope of valuable findings vs. cycles, estimated from the payout series with uncertainty bands; (2) reviewer–human agreement and score stability to validate the payout as a meaningful signal; (3) saturation detection time and false-alarm rate under change-point and sublinear-fit detectors; and (4) post-intervention area-under-payout improvement relative to no-op, which measures whether meta-actions restore marginal value per cycle. Change-point detection and sublinear modeling provide the statistical backbone for saturation timing in noisy, nonstationary conditions, while hybrid detectors mitigate failure modes in long-memory regimes [r6, wenyu2020, aminikhanghahi2017]. The meta-controller formalism ensures that interventions are selected with long-horizon objectives and a "do nothing" default, reducing unnecessary churn, and hybrid bandit–Bayesian optimization offers strong anytime performance for discrete intervention selection given low-dimensional state summaries [r34, r69, cho2412, falkner2018]. Together, these elements operationalize reviewer feedback as a market signal and show how meta-controlled interventions—novelty boosts, meta-population resets, non-elitist acceptance, iterated local search, representation refresh, re-annealing, and principled posterior discounting—can be scheduled to delay or break plateaus and sustain discovery yield under fixed cycle budgets [r28, r60, r70, r74, conti2018, ingber1989].

Trajectory Sources

Trajectory r6: The literature confirms that the saturation point can be effectively identified either by applying advanced change point detection algorithms or by fitting sublinear growth models such as logistic or power-law curves to the cumulative data (wenyu2020 pages 10-13, aminikhanghahi2017as...

Trajectory r28: Yes—the surveyed literature in organizational learning and evolutionary/complex-systems provides concrete, automatable strategies for overcoming local optima/plateaus that map directly to interventions in an AI discovery system.

Trajectory r34: Yes—the meta-RL and HRL literature provides explicit state, action, and reward formalisms to define an intervention-triggering meta-controller with time-series state features, a discrete intervention set including a 'do nothing' option, and long-horizon, post-intervention reward objectives. (cho2412...

Trajectory r60: Yes—the literature provides explicit algorithms and pseudocode that modulate Thompson-Sampling posteriors to control exploration by resetting, discounting, or windowing sufficient statistics, often triggered by external change-detection signals. (ghatak2022...

Trajectory r69: The literature supports that hybrid bandit–Bayesian optimization methods such as BOHB and BOASF are indeed well-suited for a meta-controller that selects the best discrete intervention based on low-dimensional performance time-series signals (falkner2018 pages 12-13, falkner2018bohbrobu...

Trajectory r70: Yes—Gaussian DLMs and their multivariate extensions encode an explicit state-evolution step with discount-factor parameterizations (δ) that deterministically map the previous posterior covariance into the next prior via $\mathrm{R_t} = \mathrm{G_t}$ C_{t−1} $\mathrm{G_t}$ $/\delta$ and $\mathrm{W_t} = \mathrm{G_t}$ C_{t−1} $\mathrm{G_t}$ (δ^{−1} − 1), providing a pri...

Trajectory r74: The literature supports that a meta-controller triggering a re-annealing schedule—a sharp, temporary increase in exploration intensity (via increased "temperature") followed by a gradual decay—can indeed force a low-level bandit to escape a local optimum (battiti2009 pages 31-32, in...