THE NOVELTY TRAP: WHEN REPETITION HELPS LLM RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Novelty is a common objective in automated research, but forbidding repetition can reduce the value of what agents discover. We demonstrate this *novelty trap* on the Koijen–Levy earnings-announcement benchmark. An LLM proposes economic hypotheses, a second model measures them from event summaries, and hypothesis sets fitted in one quarter are evaluated in two later quarters. At matched proposal budgets, independent restarts improve incremental R^2 by 23% and 19% relative to a sequential no-repeat prompt (0.63 and 0.57 percentage points). The advantage holds on all six seeds in each quarter, including evaluation of the same runs on historical labels withheld until the analysis protocol was fixed. A memory-preserving prompt that permits repetition also improves value on every seed. We characterise the exploration–measurement tradeoff with a model separating errors in how hypotheses represent constructs from errors in their annotation. Under its assumptions, an informative prior and noisy measurements create a region in which repeated formulations of strong ideas outperform broader coverage. The model also predicts exactly when additional annotation can reverse that ranking. Planted-truth experiments illustrate the value of exploration when prior information is weak. Explanatory value and discovery counts also diverge: counting significant coefficients in joint regressions reverses the policy ranking. The design implication is to broaden construct coverage while preserving opportunities to refine useful mechanisms, and to evaluate both by the joint information they add.

1 INTRODUCTION

Research agents must decide whether to develop familiar ideas or search for new ones. Many systems encourage the latter: prompts request new hypotheses, archives reward diversity, and filters remove near-duplicates (Lu et al., 2024; Lange et al., 2026; Romera-Paredes et al., 2024; Mouret & Clune, 2015). This has a clear rationale. Novelty search can escape deceptive objectives (Lehman & Stanley, 2011), and LLM-generated ideas are often repetitive (Si et al., 2025). Yet a familiar idea need not be exhausted. In empirical research, different formulations of the same mechanism can provide different, noisy measurements of it. Rejecting them may discard information that an unfamiliar but weak hypothesis cannot replace.

Consider earnings announcements. An agent may first propose that management guidance helps explain the market response. A later formulation might distinguish a raised forecast from a change in its uncertainty or composition. These hypotheses overlap semantically, but their measurements need not be redundant. If the model already assigns high probability to useful mechanisms, a no-repeat instruction can push it toward weaker ones. We call the resulting loss in the joint value of its hypotheses the *novelty trap*. This is a claim about a set of measured hypotheses, rather than the originality or quality of its best member.

We test this possibility on the Koijen–Levy benchmark (Koijen & Levy, 2026). Agents propose text-based conditions that may explain earnings-announcement returns beyond the earnings surprise. An annotator measures those conditions on event summaries; a model fitted to the resulting features in 2025Q4 is evaluated in 2026Q1 and 2026Q2. This setting separates proposal quality from measurement reliability and provides common outcomes against which to compare research policies. We use models with provider-reported knowledge cutoffs before the events and fix the second test quarter's

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.

analysis protocol before accessing its labels. Section 3.1 details this retrospective evaluation and its information controls.

Independent restarts produce more valuable hypothesis sets than a sequential no-repeat prompt on six of six seeds in each test quarter. The gains are 0.63 and 0.57 percentage points of incremental R^2, or 23% and 19% relative to the no-repeat arm. A control that retains memory but permits repetition also improves on the no-repeat prompt on all six seeds. A sensitivity analysis with the entire prediction head frozen supports the same direction of effect.

Our theoretical account separates two sources of measurement error: how well a hypothesis's wording represents a construct, and how reliably an annotator scores that wording. Re-annotating a fixed text addresses the second source; proposing another formulation can address both. Under explicit independence and linearity assumptions, we derive a threshold at which exploring a new construct becomes more valuable than measuring a known one again. We also characterise a boundary between informative-prior, noisy-measurement settings that favour repetition and settings that favour coverage. Developed after the main finding, the model turns it into a testable account of when novelty helps and when it hurts.

The paper makes three contributions:

- **Evidence of a novelty trap in research agents.** Matched proposal budgets, six fresh seeds, a memory-preserving control and evaluation on withheld historical labels show when a no-repeat instruction reduces the joint value of discovered hypotheses (Section 3.2).
- **A measurement-based account of repetition.** An exclusion identity distinguishes semantic similarity from informational redundancy. A two-layer reliability model yields a novelty threshold and a phase boundary between repetition and coverage (Section 4).
- **An evaluation that separates research objectives.** Explanatory value and joint-regression significance counts rank the same policies differently. Planted-truth experiments further separate discovery gains from selection error and the cost of additional evidence (Sections 5 and 3.2).

2 Related work

Research agents and self-improvement. Prompt optimisation, evolving playbooks and score-conditioned search improve task artifacts (Agrawal et al., 2026; Zhang et al., 2026a; Yang et al., 2024). Self-modifying and evolutionary systems extend this approach to code and scientific research (Zelikman et al., 2024; Zhang et al., 2025; Lange et al., 2026; Romera-Paredes et al., 2024; Srikanth et al., 2026; Guo et al., 2026; Li et al., 2026; Pan et al., 2026). Such gains need not persist across seeds, compound over iterations, or exceed strong zero-shot baselines (Xu et al., 2026; Ning et al., 2026; Oved et al., 2026; Wang et al., 2026; Zhang et al., 2026b). We isolate one consequential design choice within these systems: whether to discourage repetition.

Novelty, diversity and repeated sampling. Novelty search and quality-diversity methods reward coverage (Lehman & Stanley, 2011; Mouret & Clune, 2015); repeated sampling and sequential revision allocate inference compute differently (Brown et al., 2024; Snell et al., 2025; Zeng et al., 2025). Studies of LLM ideation document duplication and fixation (Si et al., 2025; Deng et al., 2026), and a tradeoff between originality and quality under novelty instructions (Padmakumar et al., 2026). Closest to our setting, Sen Gupta et al. (2026) find benefits from a novelty instruction in both significant-hypothesis counts and out-of-domain accuracy, while retaining an inner loop for local refinement. We obtain a different ranking for joint explanatory value under a no-repeat prompt without score feedback. ExperiGen also varies prior strength and annotation noise. We study the conditional value of repeated formulations and show that discovery counts and joint value can disagree within the same setting. The allocation problem connects to the bandwidth–fidelity tradeoff in measurement (Cronbach & Gleser, 1965) and resampling under noisy optimisation (Jin & Branke, 2005). Classical generalizability theory separates item and rater variance (Huysamen, 1990); here those components inform a research-policy allocation decision. Correlated LLM errors and homogeneous outputs limit the gains from repetition (Kim et al., 2025; Jiang et al., 2025; Clemen & Winkler, 1985).

Finite evidence and hypothesis generation. Selection bias and adaptive reuse of validation data can distort measured progress (Smith & Winkler, 2006; Ng, 1997; Cawley & Talbot, 2010; Recht et al., 2019; Dwork et al., 2015; Andrews et al., 2024; Gao et al., 2023). Agent studies find benefits from reserved final selection (Toledo et al., 2025; Hambardzumyan et al., 2026), or from validation-free selection under tight budgets (Borthwick et al., 2026). We distinguish agent-visible search evidence, a hidden selection set and later test quarters, following the separation used by AIRA2. Our task builds on LLM hypothesis and feature generation (Zhou et al., 2024; Zhong et al., 2023; Hollmann et al., 2023; Movva et al., 2025; Ludwig & Mullainathan, 2024; Manning et al., 2024). Planted-truth environments make discovery measurable (Jansen et al., 2024; Gandhi et al., 2025; Liu et al., 2025); forecasting benchmarks emphasise evaluation after model cutoffs (Halawi et al., 2024; Karger et al., 2025). Finance adds a long-standing concern with multiple testing in discovery (Harvey et al., 2016; Jensen et al., 2023) and evidence that text-derived signals predict returns (Chen et al., 2022).

3 Research agents on a real benchmark

3.1 Design

Task and data. The Koijen–Levy benchmark measures how much a prediction adds to the R^2 of an OLS regression of within-quarter abnormal-return ranks on earnings-surprise ranks (Koijen & Levy, 2026). We use the benchmark's 2025Q4 release for search and fitting (1,844 announcements, 9 October–14 November 2025). Its 1,380 search events provide the only outcomes available to the agent. Another 464 events, on every fourth label date, form a hidden selection holdout. This holdout selects single hypotheses in the relevant analyses; both subsets train the models that combine hypothesis sets. The test quarters contain 2,307 announcements on 59 dates in 2026Q1, and 1,996 announcements on 59 dates in 2026Q2.

All experiments were run retrospectively. The authors had analysed 2026Q1 earlier in the project. For Study 4, 2026Q2 labels were withheld from analysis until 26 September 2026, 04:17:15 UTC. The evaluation protocol and threshold-policy arm were frozen before unblinding. An amendment added invite-repeat and its tests after facts had been opened for annotation, but before labels were unblinded. Later benchmarking used these already-exposed quarters. Appendix B records the study-specific timing, code changes and deviations; these are local protocol freezes, not a public prospective registration.

Models and information access. The proposer is `gpt-5-2025-08-07` with medium reasoning; the annotator is `gpt-5-mini-2025-08-07` with minimal reasoning. Their provider-reported knowledge cutoffs are 30 September 2024 and 31 May 2024, respectively, before the first event on 9 October 2025. Neither uses tools. Both read the benchmark's archived fact summaries, ten facts per event. The archive does not identify the model or generation time of these summaries, so the model cutoffs do not rule out contamination through the inputs. Archived baseline predictions and published benchmark systems also use summaries, but not necessarily the same archived versions.

Agents and budgets. Each proposal contains a condition testable from the facts, a predicted sign and a proposed mechanism. The annotator scores the condition from -2 (clearly the opposite) to $+2$ (clearly present), with four events and ten conditions per call. A hypothesis's utility is its partial correlation with return rank, controlling for surprise rank, multiplied by its stated sign. Each run has six proposer calls, each requesting ten hypotheses. All arms receive the task, broad mechanism categories and the same 30 labelled search examples. From round 2 onward, the appended context differs:

- *No-repeat*: earlier conditions and an instruction not to repeat them. *Restart*: an independent call with no earlier conditions.
- *Raw*: earlier conditions and in-sample utilities, sorted best first. *Split*: earlier conditions and split-sample verdicts. Both request new hypotheses.
- *Invite-repeat*: earlier conditions and permission to re-propose important mechanisms in the same or different words.

Table 1: Search-allocation contrasts in hypothesis-set value (points of R^2; mean over Study 2's six seeds, seeds with a positive difference, paired p) in the next quarter and the sealed quarter. [a]Prespecified (Study 2; restart − no-repeat two-sided, per hypothesis one-sided); [b]prespecified (Study 4; one-sided except invite-repeat contrasts, two-sided); [c]exploratory, two-sided. 2026Q2 re-scores the same runs. Last row: per-hypothesis mean utility (the nine seeds of Studies 1 and 2 in 2026Q2).

	2026Q1			2026Q2 (sealed)		
Contrast	mean	pos.	p	mean	pos.	p
restart − no-repeat (memory, no repeat)	0.63^a	6/6	0.007	0.57^b	6/6	0.003
invite-repeat (may repeat) − no-repeat	0.47^c	6/6	0.073	0.42^b	6/6	0.002
restart − invite-repeat	0.17^c	4/6	0.186	0.15^b	4/6	0.230
restart − raw	0.52^c	6/6	0.003	0.16^b	4/6	0.152
threshold policy − restart	-0.26^c	1/6	0.248	-0.25^b	1/6	0.940
raw − no-repeat, per hypothesis	0.0047^a	5/6	0.022	0.0031^b	6/9	0.107

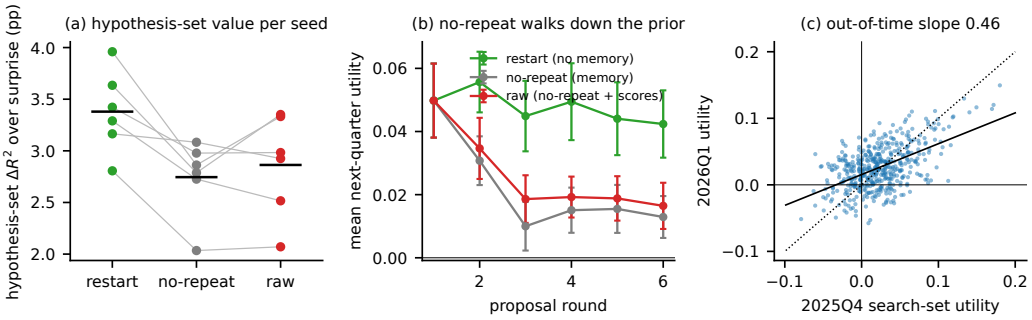

Figure 1: The novelty trap (Study 2, 2026Q1). (a) Hypothesis-set value per fresh seed (lines join a seed). (b) Mean next-quarter utility of the hypotheses proposed in each round (round 1 is shared): later no-repeat proposals have lower utility, whereas restart proposals remain stronger. (c) Search-set versus next-quarter utility of 480 Study-1 hypotheses, with least-squares fit and identity line.

- *Threshold*: each round either proposes through a restart or re-annotates the current top ten, according to estimated reliability and utility on 2025Q4.

Matched proposal allowances need not imply identical token costs; the appendix reports resource use.

Studies and outcomes. Study 1 compares no-repeat, raw and split on three seeds. Study 2, specified after Study 1, compares restart, no-repeat and raw on six fresh seeds. Studies 3a and 3b repeat annotations. Study 4 evaluates the existing runs on withheld 2026Q2 labels and adds invite-repeat and threshold on Study 2's seeds. A hypothesis set's score combines its features using ridge weights (penalty 1000) fitted on 2025Q4. Its primary value is the score's incremental R^2 over surprise in the test-quarter OLS. Because that OLS refits the score's coefficient, the metric measures out-of-time explanatory value, rather than a fully frozen forecast. We separately report a frozen-head sensitivity analysis. Seed-paired tests measure variation across six runs conditional on the evaluated events; date-resampling analyses address a different source of uncertainty.

3.2 The novelty trap

Restarts beat the no-repeat instruction in both test quarters. Independent restarts improve hypothesis-set value by 0.63 percentage points in 2026Q1 and 0.57 points in 2026Q2, at the same proposal budget (Table 1). Every one of the six fresh seeds favours restarts in the first quarter (prespecified, two-sided $p = 0.007$; exact sign-flip $p = 0.031$), and the same six runs all favour restarts on the withheld quarter (prespecified, one-sided $p = 0.003$). The history-based no-repeat policy yields a less valuable set despite access to earlier ideas.

Permitting repetition preserves the benefits of memory. Invite-repeat retains the agent's history and allows it to re-propose important mechanisms. It improves on no-repeat on every seed: 0.42 points in 2026Q2 (two-sided $p = 0.002$) and 0.47 in 2026Q1 (exploratory, $p = 0.073$). The remaining gap to restarts is not significant (0.15 points, four of six seeds, 95% interval $[-0.14, 0.44]$). Thus discarding memory is not necessary to improve on no-repeat. The prompt changes both the no-repeat instruction and the framing of the history; a neutral same-history control would separate these effects.

Robustness across estimation choices. The 2026Q1 restart advantage persists across ridge penalties from 0 to 10^5 (0.50–0.63 points, at least 5 of 6 seeds), and equal, correlation-based or cross-validated weights (0.49–0.96). Removing near-duplicates leaves 0.34 points (six of six seeds); matching the hypothesis count leaves 0.43. Standardising features on the search quarter gives 1.00 and 1.04 points in the two test quarters; freezing the entire prediction head gives 0.70 and 0.58 (six of six seeds in each case). Dropping the one 2026Q2 summary that mentions an intraday price reaction leaves 0.58 points. An exploratory bootstrap jointly resampling run blocks and evaluation dates gives 95% intervals of [0.123, 1.195] and [0.057, 1.067] points, conditional on the fitted Q4 models (Appendix B.2). Date-only resampling gives [0.29, 1.00] for Q1. The distinct, prespecified comparison of arm-pooled sets is less conclusive (0.81, [-0.12, 1.66]).

The benefit depends on hypothesis content and the evaluation objective. Excluding guidance, outlook, forecast and target hypotheses leaves 0.84 points in 2026Q1 ([0.17, 1.52]), but 0.02 in 2026Q2 ([−0.43, 0.48]). The second-quarter advantage is therefore concentrated in guidance-related mechanisms, which every run proposed early. Counting individually significant hypotheses reverses the ranking: no-repeat yields 1.67 and 1.50 per run, versus 0.33 and 0.50 for restarts (joint regression, Bonferroni correction, date-clustered inference). More measurements of the same construct can improve their joint value while reducing the power of individual coefficients (Appendix A). This within-experiment reversal helps explain why evaluations based on discovery counts and those based on joint value can reach different conclusions about novelty (Sen Gupta et al., 2026).

Evidence consistent with an informative prior. The results suggest that new measurements of strong mechanisms can add more information than proposals from weaker parts of the prior. Two patterns support this account, formalised in Section 4. Later no-repeat proposals have lower utility (consistent with Proposition 3c): mean utility falls from 0.050 in the shared first round to 0.017 (no-repeat) and 0.022 (raw) in later rounds, while restarts stay at 0.047 (Figure 1b). The prior is measurably informative (Proposition 3d): within restart runs, how often a mechanism recurs—a hypothesis-level analogue of self-consistency (Wang et al., 2023)—predicts its next-quarter utility (Spearman 0.51) and its utility on the never-shown holdout (0.16–0.37 across seeds, all positive). One repeat annotation of each run's best first-round hypothesis adds 0.14–0.31 points, whereas a later new hypothesis adds 0.001–0.032 on average and at most 6% of them beat the re-annotation; only 10–16% of later sequential proposals clear the re-annotation threshold $\tau(0.70) = 0.42$, against 37% for restarts.

Re-annotation did not close the gap. Annotations are noisy (test–retest 0.70 in identical batches, 0.49 regrouped), and averaging three annotations raises the mean utility of the selected hypotheses by 12.9% (Study 3a; Spearman–Brown implies 11.7%). Study 3b predicted that raising annotation reliability to about 0.875 for every Study-2 hypothesis would shrink the restart advantage. This prespecified prediction failed (0.00 points, $[-0.29, 0.30]$). This result motivates separating annotation noise from the quality of the underlying operationalisation (Section 4). Post hoc calibration does not distinguish the one-layer and two-layer accounts once the empirical coverage component is included (Appendix A.1). The threshold policy, which re-measured by re-annotation, did not beat restarts (-0.25 points in the sealed quarter, one of six seeds positive).

Feedback changes the comparison. Raw score feedback increases the next-quarter utility of individual hypotheses in Study 2 (0.0047, 5 of 6 seeds, prespecified one-sided $p = 0.022$). The corresponding gain is not significant in 2026Q2 (nine seeds, one-sided $p = 0.107$). Restarts beat raw feedback on all six seeds in Q1, but the prespecified Q2 comparison does not establish a difference (0.16 points, $p = 0.152$). Raw feedback and invite-repeat have similar Q2 values (3.36 versus 3.37 points). On Study 2's seeds, feedback improves the point estimate of set value over no-repeat

in both quarters, without a significant difference. The prediction from Study 1 that it would reduce set value was not supported (Appendix B).

3.3 HYPOTHESIS SETS EXPLAIN LATER-QUARTER RETURNS

The selected Study-1 hypotheses, combined with 2025Q4 weights, add 3.96 and 3.67 percentage points of R^2 over earnings surprise in the two test quarters. In 2026Q1, the gain remains 3.79 with every coefficient frozen in advance. The hypotheses also add 0.74 and 0.20 points beyond both archived LLM baselines, with significant prespecified increments in both quarters (one-sided $t = 1.9$ in Q2; Appendix Table 3). They add 2.69 points beyond supervised TF-IDF in Q1 and remain informative beyond a direct GPT-5 forecast. The value of explicit hypothesis measurement is therefore complementary to direct prediction.

Of the 480 unique condition texts, 147 have positive next-quarter utility in their stated direction at a Benjamini–Hochberg rate of 10%. Among the 86 with search-set $t > 2$, 96.5% retain their sign (versus 75% of the rest), and 55 pass the same test. As expected under noisy selection, utility attenuates out of time: its regression slope on search-set utility is 0.46 in Q1 and 0.48 in Q2, versus 0.73 on the hidden selection holdout. Utilities correlate 0.58 across test quarters. Single-winner selection does not distinguish the policies: search-best, holdout and pooled top-five rules select the same first-round guidance-raise hypothesis in every run.

What the agents found. The agents translate economic constructs into specific, testable conditions (Appendix Table 4): guidance issued with earnings, cash conversion, deleveraging, operating leverage, fixed-cost absorption, adverse mix, channel inventory and platform changes. These are familiar economic ideas—guidance bundled with earnings is a well-known confound of the surprise response (Rogers & Van Buskirk, 2013)—but their explicit measurement adds information to the benchmark's LLM predictions. The content extends beyond guidance: 63 selected hypotheses excluding guidance, outlooks, forecasts and targets add 0.50 points beyond both baselines in Q1 ($t = 3.3$).

3.4 COMPARISON WITH REPORTED BENCHMARK SYSTEMS

Study 5 evaluates prediction systems on the benchmark's own split (Appendix C). Our best system without archived predictions scores 12.2% under the benchmark's convention. Published reference scores are 12.8% for GEPA, 14.1% for Opus 4.5 zero-shot and 17.1% for Opus with ACE. These comparisons cross model families and input provenance, and are not controlled estimates of an algorithm's advantage or official leaderboard submissions.

Within our OpenAI systems, playbooks produce predictions strongly correlated with zero-shot forecasts (0.90–0.93), and the tested-mechanism playbook does not improve on the ACE-style control. Averaging six models yields modest gains, consistent with their correlated predictions. An explicitly measured hypothesis score adds information to the ensemble in all three samples, significantly on the benchmark sample. These secondary comparisons use data exposed in earlier studies; Appendix C reports the designs, comparisons and uncertainty.

4 RESEARCH VALUE UNDER NOISY MEASUREMENT

Discovery and selection. For a candidate pool $\mathcal{C}$ with selected candidate $\hat{f}$, discovery quality $D = \max_{f \in \mathcal{C}} U(f)$ and selection regret $S = D - U(\hat{f})$ give $U(\hat{f}) = D - S$; both terms are observable when truth is known (Section 5).

When can a novelty constraint hurt? Let $V(S)$ be the best predictive value obtainable from a hypothesis set S, with the option to ignore any hypothesis, and $G(h \mid S) = V(S \cup \{h\}) - V(S) \geq 0$ the conditional value of a new proposal h. A novelty constraint rejects proposals in a neighbourhood $B(S)$ of what is already held.

Proposition 1 (Exclusion boundary)**.** *Let q be the unconstrained proposal law, $b = q(B)$, and, for $0 < b < 1$, μ_B and μ_A the mean of G under q inside and outside B. One unconstrained proposal is worth $\mathbb{E}_q G - \mathbb{E}_{q(\cdot \mid h \notin B)} G = b\,(\mu_B - \mu_A)$ more than one constrained proposal. Excluding identical*

observed features with no new measurement ($\mu_B = 0$) therefore never lowers value; a constraint lowers it exactly when what it excludes is conditionally more valuable than what it admits.

Semantic similarity is not informational redundancy: a new wording of a known construct, or a different mechanism in the same family, can carry high conditional value. The model below gives one concrete reason, measurement noise, and says when it dominates.

Constructs, items and two layers of reliability. A hypothesis operationalises a latent construct F_i (say, guidance raised with the release) whose utility $u_i \geq 0$ is its correlation with the outcome given the controls. The hypothesis itself is an *item*: a condition text whose true score is the construct plus wording-specific error, rated by an annotator with further error. Two reliabilities follow. *Annotation reliability* ρ_a is the test–retest correlation of the annotator on the same text; *operationalisation reliability* ρ_{op} is the share of the text's true-score variance that is the construct. One annotated item measures its construct with reliability $\rho = \rho_a \rho_{op}$; the mean of m distinct items, each annotated k times, measures it with $R(m,k) = \mathrm{SB}_m\{\rho_{op}\,\mathrm{SB}_k(\rho_a)\}$, where $\mathrm{SB}_m(r) = mr/\{1+(m-1)r\}$ (Spearman, 1910; Brown, 1910). Re-annotating a text in the same rating context is capped at ρ_{op}; re-proposing the construct in new texts is not. For orthogonal constructs, independent errors and a large-sample linear fit (assumptions (A1)–(A5), Appendix A), a hypothesis set holding m_i items of construct i has value $V = \sum_i u_i^2 R(m_i, k)$ (Lemma 2).

Prior informativeness. The proposer draws construct i with probability $p_i \propto u_i^\kappa$. $\kappa = 0$ is an uninformative prior, as in our planted tasks; large κ proposes the strongest constructs first and most often. We idealise the no-repeat instruction as sampling constructs *without replacement*, and independent restarts as sampling *with replacement*. Actual prompt compliance need not implement either distribution exactly.

Proposition 2 (Novelty threshold)**.** *Under (A1)–(A5), given a hypothesis set that holds the best construct once (one item, annotated once), a previously unrepresented construct j adds more value than re-proposing the best construct in a new text if and only if $u_j/u_1 > \tau(\rho) = \sqrt{(1-\rho)/(1+\rho)}$ with $\rho = \rho_a\rho_{op}$, and more than re-annotating the existing text if and only if $u_j/u_1 > \tau(\rho_a)$. For $0 < \rho < 1$, a re-annotation is worth a fraction $(1-\rho_a)(1+\rho)/\{(1+\rho_a)(1-\rho)\} \leq 1$ of a re-proposal, with equality only if $\rho_{op} = 1$. At $\rho = 1$, both repeat gains vanish.*

Proposition 3 (With versus without replacement)**.** *Under (A1)–(A5), with $2 \leq K < N$ and positive, nonconstant utilities u_i, let $\Delta(\kappa, \rho)$ be the expected value of K proposals (each a new item annotated once) drawn with replacement minus K drawn without. (a) If $\kappa = 0$, $\Delta < 0$ for every $\rho \in (0,1]$; if $\rho = 1$, drawing without replacement is never worse, whatever κ. (b) Δ/ρ is strictly decreasing in ρ, so there is a boundary $\rho_b(\kappa) \in [0,1]$ with $\Delta > 0$ if and only if $\rho < \rho_b(\kappa)$; $\rho_b(\kappa) > 0$ for every $\kappa > 0$, and if $u_1 > u_2$, $\rho_b(\kappa) \to (K/Q_K - 1)/(K-1)$ as $\kappa \to \infty$, where $Q_K = \sum_{i \leq K}(u_i/u_1)^2$. (c) Under an informative prior, the expected utility of the r-th new proposal decreases in r, while each restart draw keeps the first draw's expected utility. (d) The number of restarts that re-propose a construct is stochastically increasing in its utility.*

Proofs and assumptions are in Appendix A.

A testable annotation crossover. Let $b = \rho_b(\kappa)$ and suppose the initial reliability $\rho_a\rho_{op} < b$ favours repetition. Giving every item k annotations in both policies raises effective reliability to $r_k = \rho_{op}\mathrm{SB}_k(\rho_a)$. If $\rho_{op} \leq b$, no finite k crosses the boundary. If $\rho_{op} > b$, re-annotation reaches the boundary when

$$k \geq k_* = \frac{b(1-\rho_a)}{\rho_a(\rho_{op} - b)}.$$

Choose the smallest integer satisfying the inequality: equality gives a tie, and $k > k_*$ reverses the ranking. Thus the reliability ceiling determines whether re-annotation can remove the novelty trap. This prediction fixes proposal count while increasing annotation spending in both policies; it is not a fixed-dollar optimum (proof in Appendix A).

Post hoc estimates on the Koijen–Levy data give $\rho_a = 0.73$, $\rho_{op} = 0.73$ and $\hat{\kappa} = 1.1$ $[0.9, 1.4]$ (Appendix A.1). Under the illustrative profile $u_i \propto i^{-1}$, these estimates place the benchmark in the repetition-favouring region of Figure 2; alternative estimators give 0.71–0.8 under $u_i \propto i^{-1}$. The flatter $i^{-1/2}$ profile is sensitive to the choice of estimator (Appendix A.1). The corresponding

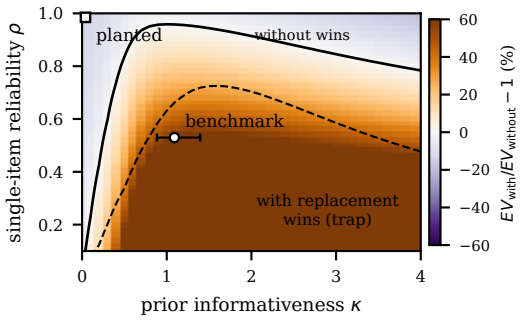

Figure 2: Predicted gain from proposing with rather than without replacement ($K = 60$ proposals over $N = 200$ constructs, $u_i \propto i^{-1}$). Solid: zero contour; dashed: zero contour for $u_i \propto i^{-1/2}$. Circle: illustrative Koijen–Levy calibration with 95% intervals conditional on estimation choices; square: planted tasks, placed by design. Appendix Figure 3 gives both profiles.

novelty thresholds are 0.55 for re-proposal and 0.40 for re-annotation. This calibration is approximate: reliability uses raw rating correlations, whereas utility conditions on the earnings surprise. The planted tasks represent an uninformative-prior setting with deterministic candidate evaluation; their placement at $\kappa = 0, \rho = 1$ is conceptual, not an estimated fit. A controlled test varying both prior informativeness and measurement reliability remains necessary.

5 Planted-truth experiments: when exploration pays

In planted signal-discovery tasks, inputs are point-in-time features on an asset×date panel, labels are $y = \mu(X) + \varepsilon$ with a planted mechanism μ (the experiments use linear, absolute-difference and maximum interaction mechanisms, plus a zero-signal null family) and noise calibrated to daily returns, and a candidate is a typed expression program with one fitted coefficient. Searchers see noisy gains on a search set, the only labels they may query; shortlisted candidates may be scored on a priced validation panel (the planted analogue of a selection holdout); only the evaluator knows the noiseless utility on an independent design (Appendix D). Here the proposer's prior carries no information about μ.

P1: feedback drives discovery. A frozen `gpt-5.6-luna` editor makes 128 proposals per trajectory over 3 tasks, 6 noise worlds and 2 policy tables. Raw search-set feedback raises selected true utility by $+0.00917$ ($[+0.00476, +0.01313]$); without feedback, utility remains flat in depth. At depth 128, the best candidate has utility 0.0234 with feedback and 0.0150 without, so discovery accounts for 92% of the gain. An independent top-eight panel does not recover the remaining selection regret ($+0.00013$, $[-0.00049, +0.00071]$). The search–truth gap grows in both arms, showing that greater optimism need not imply worse discoveries. Inference conditions on the recorded policy tables (Appendix D).

P2: the value of validation depends on precision and on the selector. Reserving 8 of 128 classical-search queries for independent selection raises utility by $+0.00050$ ($[+0.00017, +0.00084]$) over 32 fresh seeds, at equal query count but with additional validation labels. In a 64-seed follow-up, holdout-only selection loses with small panels and gains only with the largest. Pooling search and validation scores by sample size removes the small-panel losses and improves on all-search at equal or fewer row evaluations (Appendix Table 12). Precision and selection rules therefore help reconcile differing findings on the value of validation (Hambardzumyan et al., 2026; Borthwick et al., 2026).

P3: learned selectors transfer mainly by abstaining. A contextual selector trained on source tasks improves target utility over pooling, but 99% (hill-climbing) and 62% (evolutionary search) of the gain comes from a zero-signal family on which it rejects every candidate (Appendix D.3).

6 DISCUSSION AND LIMITATIONS

Implications for research agents. A search policy can produce more diverse hypotheses yet explain less of the outcome. In our benchmark, allowing repeated proposals preserves useful measurements of familiar mechanisms. In planted tasks with little prior information, feedback instead improves the quality of discovered candidates. Research systems should use novelty to broaden construct coverage while preserving local refinement of promising mechanisms. Repeat-permitted and independent-restart baselines make the cost of suppressing that refinement measurable. Novelty, joint explanatory value and individually significant discoveries answer different questions; none alone measures overall research progress.

The next methodological step is an agent that learns when to explore a new construct and when to improve an existing measurement. Our threshold policy did not outperform restarts, so the theory currently motivates that design problem rather than solves it.

Limitations. The main evidence comes from one benchmark, one proposer–annotator pair and two later quarters. The six-seed contrasts condition on those events; the prespecified date-resampling contrast of arm-pooled sets includes zero. The 2026Q2 evaluation reuses existing runs, so it tests temporal transfer rather than replication across new seeds. Its authors-withheld labels protect this comparison more strongly than the repeatedly analysed 2026Q1 data, but do not remove adaptation elsewhere in the project or unknown contamination in archived summaries. Changes in disclosure regimes may also affect transfer.

Invite-repeat changes both the prohibition and the framing of the history. It shows that memory is compatible with strong performance, but a neutral same-history prompt is needed to isolate the prohibition itself. The advantage is sensitive to hypothesis content: excluding guidance-related hypotheses removes clear support for it in 2026Q2. The post hoc theory assumes orthogonal constructs and independent errors, and its calibration does not distinguish all competing measurement models. A controlled phase-diagram experiment remains outstanding. Reported outcomes are explanatory R^2, not trading returns. The appendices document specified endpoints, exploratory analyses, deviations and resource use.

Conclusion. Repetition can be productive when research turns ideas into noisy measurements. On the Koijen–Levy benchmark, forbidding earlier ideas reduces the joint explanatory value of the resulting hypothesis sets relative to restarts and to a memory-preserving prompt that permits repetition. A familiar mechanism is not an exhausted measurement opportunity: research agents should preserve useful refinements and judge proposals by the information they add.

AI USE STATEMENT

Large language models served as experimental agents and assisted with research and writing. In the real-benchmark studies (Section 3), the proposer was `gpt-5-2025-08-07` with medium reasoning, the annotator was `gpt-5-mini-2025-08-07` with minimal reasoning, and the direct-forecast comparator was `gpt-5-2025-08-07` with low reasoning. Study 5 used `gpt-5.2-2025-12-11`, `gpt-5.1-2025-11-13`, `gpt-5-2025-08-07`, `o3-2025-04-16`, `o4-mini-2025-04-16` and `gpt-4.1-2025-04-14`. We checked served identifiers against the receipts and selected dated snapshots whose provider-reported knowledge cutoffs precede the events.

The editor in planted-truth experiment P1 was `gpt-5.6-luna` with medium reasoning; archived receipts confirm its returned identifier. The `gpt-5.6-sol` and `gpt-5.4-mini` runs produced incomplete grids (Table 8). These tasks are synthetic and do not use future financial outcomes. LLM agents also assisted with analysis code, literature search, numerical and claim checks, prose, and LaTeX. Those checks are internal and do not constitute independent human review. The authors remain responsible for the research designs, interpretations, references and final submission, including errors introduced through AI assistance.

REPRODUCIBILITY STATEMENT

For the real-benchmark studies, the supplement provides agent, annotation, prediction and analysis code; frozen protocols and amendments with hashes; generated hypotheses with search-set and out-of-time statistics; annotator ratings and predictions; and usage and cost summaries with credentials excluded. Offline analyses reproduce the primary empirical tables of Studies 1, 2, 4 and 5 from saved outputs. Exact regeneration depends on the numerical environment and complete cost-ledger inputs; absent ledger inputs should not be interpreted as zero cost. The theory calibration also uses rounded intermediate inputs and has not been independently regenerated in full. Numerical result macros are generated from result files, with new review-stage checks identified separately. The underlying event archives must be obtained from the benchmark provider and are not redistributed.

For planted-truth experiments P1–P3, the supplement provides archived trajectories, reconstructed utilities for every recorded candidate, code to regenerate P2, and the selector-transfer replay. These materials support reanalysis; they do not independently replicate the LLM runs.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html`.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024. doi: 10.1093/qje/qjad043.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1006–1014, 2015.

Andrew Borthwick, Stephen Ash, and Anthony Galczak. RoboPhD: Evolving diverse complex agents under tight evaluation budgets. *arXiv preprint arXiv:2604.04347*, 2026. URL `https://arxiv.org/abs/2604.04347`.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL `https://arxiv.org/abs/2407.21787`.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Yifei Chen, Bryan T. Kelly, and Dacheng Xiu. Expected returns and large language models. SSRN working paper 4416687, 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4416687.

Robert T. Clemen and Robert L. Winkler. Limits for the precision and value of information from dependent sources. *Operations Research*, 33(2):427–442, 1985.

Lee J. Cronbach and Goldine C. Gleser. *Psychological Tests and Personnel Decisions*. University of Illinois Press, Urbana, IL, 2nd edition, 1965.

Yuting Deng, Melanie Brucks, and Olivier Toubia. Examining and addressing barriers to diversity in LLM-generated ideas. *arXiv preprint arXiv:2602.20408*, 2026.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Kanishk Gandhi, Michael Y. Li, Lyle Goodyear, Agam Bhatia, Louise Li, Aditi Bhaskar, Mohammed Zaman, and Noah D. Goodman. BoxingGym: Benchmarking progress in automated experimental design and model discovery. *arXiv preprint arXiv:2501.01540*, 2025. URL https://arxiv.org/abs/2501.01540.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 10835–10866, 2023.

Jiacheng Guo, Suozhi Huang, Yunlong Gao, Zihao Li, Jason Ge, Xu Kuang, and Mengdi Wang. AQuA: Recursively self-improving quantitative trading research agents. *arXiv preprint arXiv:2608.12841*, 2026.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 50426–50468, 2024. doi: 10.52202/079017-1598. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a5acfd0876c940d81619c1dc60e7748-Abstract-Conference.html.

Karen Hambardzumyan, Nicolas Baldwin, Edan Toledo, Rishi Hazra, Michael Kuchnik, Bassel Al Omari, Thomas Simon Foster, Anton Protopopov, Jean-Christophe Gagnon-Audet, Ishita Mediratta, Kelvin Niu, Michael Shvartsman, Alisia Lupidi, Alexis Audran-Reiss, Parth Pathak, Tatiana Shavrina, Despoina Magka, Hela Momand, Derek Dunfield, Nicola Cancedda, Pontus Stenetorp, Carole-Jean Wu, Jakob Nicolaus Foerster, Yoram Bachrach, and Martin Josifoski. AIRA_2: Overcoming bottlenecks in AI research agents. *arXiv preprint arXiv:2603.26499*, 2026. URL https://arxiv.org/abs/2603.26499.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ...and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.

Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing CAAFE for context-aware automated feature engineering. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 44753–44775, 2023. doi: 10.52202/075280-1938. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c2df4c35cdbee764ebb9e9d0acd5197-Abstract-Conference.html.

G. K. Huysamen. The application of generalizability theory to the reliability of ratings. *South African Journal of Psychology*, 20(3):200–205, 1990. doi: 10.1177/008124639002000310.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, pp. 10088–10116, 2024. doi: 10.52202/079017-0324. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13836f251823945316ae067350a5c366-Abstract-Datasets_and_Benchmarks_Track.html.

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518, 2023. doi: 10.1111/jofi.13249.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 90705–90774, 2025.

Yaochu Jin and Jürgen Branke. Evolutionary optimization in uncertain environments—a survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, 2005.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea74e45a229dac70b5b63b28d8934db6-Abstract-Conference.html. arXiv:2409.19839.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 30038–30066, 2025.

Ralph S. J. Koijen and Bradford Levy. Assessing the benefits of optimized agentic AI systems for asset pricing. SSRN working paper 6474601, June 2026. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6474601. June 30, 2026 version.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. Shinkaevolve: Towards open-ended and sample-efficient program evolution. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7886b9bafe76c52fd568db10ff9772df-Abstract-Conference.html.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011. doi: 10.1162/EVCO_a_00025.

Zongqian Li, Yaoyiran Li, Yaohui Guo, Ming Zhang, Nigel Collier, and Eugene Ie. AutoScientist-Quant: Self-evolving coding agents for automatic research in quantitative investment. *arXiv preprint arXiv:2608.28632*, 2026.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. HypoBench: Towards systematic and principled benchmarking for hypothesis generation. *arXiv preprint arXiv:2504.11524*, 2025. URL https://arxiv.org/abs/2504.11524.

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.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024. doi: 10.1093/qje/qjad055.

Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientist and subjects. NBER Working Paper 32381, National Bureau of Economic Research, 2024. arXiv:2404.11794.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015. URL https://arxiv.org/abs/1504.04909.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 44997–45023, 2025. URL https://proceedings.mlr.press/v267/movva25a.html. arXiv:2502.04382.

Andrew Y. Ng. Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pp. 245–253, 1997. URL https://mlanthology.org/icml/1997/ng1997icml-preventing/.

Jingjie Ning, Shanshan Zhong, Xiaochuan Li, and Ji Zeng. Scores alone do not prove discovery: The discovery certification protocol for auditing AI research agents. *arXiv preprint arXiv:2609.09219*, 2026. URL https://arxiv.org/abs/2609.09219.

Tal Oved, Roi Pony, Oshri Naparstek, and Udi Barzelay. Evolution or illusion? Rethinking evaluation in LLM evolutionary search. *arXiv preprint arXiv:2609.19799*, 2026. URL https://arxiv.org/abs/2609.19799.

Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Measuring LLM novelty as the frontier of original and high-quality output. In *International Conference on Learning Representations (ICLR)*, 2026.

Yingjian Pan, Xiaowei Ding, and Kay Giesecke. Agentic empirical asset pricing: Methodological foundations. *arXiv preprint arXiv:2609.00731v1*, 2026. URL https://arxiv.org/abs/2609.00731v1. Version 1, 1 September 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Jonathan L. Rogers and Andrew Van Buskirk. Bundled forecasts in empirical accounting research. *Journal of Accounting and Economics*, 55(1):43–65, 2013. doi: 10.1016/j.jacceco.2012.06.001.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Jishu Sen Gupta, Harini SI, Somesh Kumar Singh, Syed Mohamad Tawseeq, Yaman Kumar Singla, David Doermann, Rajiv Ratn Shah, and Balaji Krishnamurthy. Accelerating social science research via agentic hypothesization and experimentation. *arXiv preprint arXiv:2602.07983*, 2026.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html. arXiv:2409.04109.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html. arXiv:2408.03314.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Dhruv Srikanth, Bingchen Zhao, Dixing Xu, Yuxiang Wu, and Zhengyao Jiang. Recursive self-improvement of AI research agents, September 2026. URL https://arxiv.org/abs/2609.26457v1. Preprint, version 1, September 22, 2026.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, Andrei Lupu, Roberta Raileanu, Tatiana Shavrina, Kelvin Niu, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Shagun Sodhani, Alexander Miller, Abhishek Charnalia, Derek Dunfield, Carole-Jean Wu, Pontus Stenetorp, Nicola Cancedda, Jakob Foerster, and Yoram Bachrach. AI research agents for machine learning: Search, exploration, and generalization in MLE-bench. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/328b81881da145412f2bc56c998dfb6a-Abstract-Conference.html. arXiv:2507.02554.

Wenxiao Wang, Priyatham Kattakinda, and Soheil Feizi. Do agent optimizers compound? A continual-learning evaluation on Terminal-Bench 2.0. *arXiv preprint arXiv:2607.14004*, 2026. URL https://arxiv.org/abs/2607.14004.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking. *arXiv preprint arXiv:2605.05973*, 2026. URL https://arxiv.org/abs/2605.05973v1.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html. arXiv:2309.03409.

Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (STOP): Recursively self-improving code generation. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2310.02304.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4651–4665, 2025.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y. Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a94ff6f922d995d7d3f4ebf4143e442-Abstract-Conference.html. arXiv:2510.04618.

Xing Zhang, Guanghui Wang, Yanwei Cui, Wei Qiu, Ziyuan Li, Bing Zhu, and Peiyang He. Prompt optimization is a coin flip: Diagnosing when it helps in compound AI systems. *arXiv preprint arXiv:2604.14585*, 2026b. URL https://arxiv.org/abs/2604.14585. CTB Workshop at ICML 2026.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 40204–40237, 2023. doi: 10.52202/075280-1748. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7e810b2c75d69be186cadd2fe3febeab-Abstract-Conference.html.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, Miami, FL, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://aclanthology.org/2024.nlp4science-1.10/. arXiv:2404.04326.

A A TWO-LAYER RELIABILITY MODEL OF SEARCH ALLOCATION

Setup. Let y be the standardised outcome after removing the controls. Constructs $F_1, \dots, F_N$ are standardised latent variables with utilities $u_i = \operatorname{corr}(F_i, y) \geq 0$ (the sign is fixed by the stated direction), ordered $u_1 \geq \cdots \geq u_N$ and not all equal. An *item* t of construct i has true score $T_t = F_i + \delta_t$. Its r-th annotation is $X_{t,r} = T_t + \varepsilon_{t,r}$. The operationalisation reliability is $\rho_{op} = \operatorname{corr}(T_t, F_i)^2$ and the annotation reliability is $\rho_a = \operatorname{corr}(X_{t,r}, X_{t,r'})$, $r \neq r'$. The item-specific error δ_t includes the effects of wording and any rating context held fixed across re-annotations, such as the annotation batch.

The model assumes: (A1) the constructs are mutually uncorrelated; (A2) items are parallel, with common ρ_{op} and ρ_a; (A3) the δ_t and $\varepsilon_{t,r}$ are mutually independent and independent of all F; (A4) $y = \sum_i u_i F_i + e$, with e independent of every δ and ε; (A5) the value of a hypothesis set $\mathcal{H}$ is the population R^2 of the best linear combination of its features, $V(\mathcal{H}) = \max_w \operatorname{corr}(w^\top X_{\mathcal{H}}, y)^2$. A single annotation of a single item measures its construct with reliability $\rho = \operatorname{corr}(X_{t,r}, F_i)^2 = \rho_a \rho_{op}$.

Lemma 1 (Two-layer Spearman–Brown)**.** *The mean of m distinct items of one construct, each annotated k times, has construct reliability $R(m,k) = \mathrm{SB}_m\{\rho_{op}\,\mathrm{SB}_k(\rho_a)\}$. Hence $R(m,1) = \mathrm{SB}_m(\rho)$, $R(1,k) \uparrow \rho_{op}$ as $k \to \infty$, and $R(m,k) \to 1$ as $m \to \infty$.*

Proof. Set $\operatorname{Var} F = 1$. Then $\operatorname{Var} T = 1/\rho_{op}$ and $\operatorname{Var} \varepsilon = (1-\rho_a)/(\rho_a\rho_{op})$. The mean of k annotations of one item has covariance 1 with F and variance $1/\rho_{op} + (1-\rho_a)/(k\rho_a\rho_{op}) = 1/\{\rho_{op}\mathrm{SB}_k(\rho_a)\}$. It is therefore an item with reliability $r_k = \rho_{op}\mathrm{SB}_k(\rho_a)$. By (A3), the m such items are parallel with independent errors, so their mean has variance $1 + (1/r_k - 1)/m$, and its reliability is $\mathrm{SB}_m(r_k)$. □

Lemma 2 (Additivity)**.** *Under (A1)–(A5), if $\mathcal{H}$ holds m_i items of construct i, each annotated k times, then $V(\mathcal{H}) = \sum_i u_i^2 R(m_i, k)$. The best linear combination weights the items of each construct equally, so fitting every item gives the same value as fitting the construct means.*

Proof. By (A1), (A3) and (A4), the feature covariance matrix is block-diagonal by construct, and $\operatorname{cov}(y, X_t) = u_{i(t)} \operatorname{cov}(F_{i(t)}, X_t)$. The projection of y on the features is therefore the sum of the projections on each block, and the explained variances add. Within block i the features are exchangeable and have a common covariance with y, so the projection puts equal weight on each and equals the projection on their mean $\bar{X}_i$. It explains $\operatorname{cov}(y, \bar{X}_i)^2 / \operatorname{Var} \bar{X}_i = u_i^2 R(m_i, k)$, because $\operatorname{cov}(F_i, \bar{X}_i) = 1$ and $R = 1/\operatorname{Var} \bar{X}_i$. □

If items have different reliabilities r_j, the same argument gives $R_i = 1 - 1/\{1 + \sum_j r_j/(1-r_j)\}$, with weights proportional to $1/\operatorname{Var}(\delta_j + \bar{\varepsilon}_j)$.

Proof of Proposition 2. Start with one item of construct 1, annotated once. By Lemma 2, adding a previously unrepresented construct j gains $u_j^2\rho$. Re-proposing construct 1 in a new text gains $u_1^2\{\mathrm{SB}_2(\rho) - \rho\} = u_1^2\rho(1-\rho)/(1+\rho)$. Re-annotating the existing text gains $u_1^2\rho_{op}\{\mathrm{SB}_2(\rho_a) - \rho_a\} = u_1^2\rho(1-\rho_a)/(1+\rho_a)$. Comparing these gains gives both thresholds. For $0 < \rho < 1$, the ratio of the re-annotation gain to the re-proposal gain is at most one if and only if $(1-\rho_a)(1+\rho) \leq (1+\rho_a)(1-\rho)$, that is, if and only if $\rho \leq \rho_a$, with equality only when $\rho_{op} = 1$. At $\rho = 1$, both repeat gains vanish and their ratio is undefined. □

Two corollaries. First, at the same annotation budget, m distinct items are at least as reliable as m annotations of one item: $\mathrm{SB}_m(\rho_a\rho_{op}) \geq \rho_{op}\mathrm{SB}_m(\rho_a)$ if and only if $1 + (m-1)\rho_a \geq 1 + (m-1)\rho_a\rho_{op}$. Second, an item's reliability is bounded above by ρ_{op} regardless of the number of

re-annotations. For $\rho_{op} < 1$, distinct items reach that bound when $m \geq m^* = (1-\rho)/\{\rho_a(1-\rho_{op})\}$, and exceed it when the inequality is strict. On the Koijen–Levy estimates below, $m^* = 2.4$: three single-annotated items reach $R = 0.77$, while three annotations of one item reach 0.65 and infinitely many reach only 0.73. A re-annotation is worth 0.51 of a re-proposal. The one-layer threshold $\tau(0.70) = 0.42$ used in Section 3.2 rests on a same-text retest correlation. It is therefore the threshold for re-annotation, $\tau(\rho_a)$; the threshold that governs re-proposal is $\tau(\rho_a\rho_{op}) = 0.55$.

Corollary: when additional annotation reverses the ranking. Fix the proposal budget, prior and utilities, and let $b = \rho_b(\kappa)$ be the boundary in Proposition 3. Suppose $0 < \rho_a < 1$ and the initial item reliability satisfies $\rho_a\rho_{op} < b$. Annotating every proposed item k times in both arms replaces its reliability by $r_k = \rho_{op}k\rho_a/\{1+(k-1)\rho_a\}$. If $\rho_{op} \leq b$, no finite k reverses the initial ranking; equality permits only asymptotic convergence to the boundary. If $\rho_{op} > b$, the two policies tie or favour without-replacement sampling exactly when

$$k \geq k^* = \frac{b(1-\rho_a)}{\rho_a(\rho_{op}-b)}. \tag{1}$$

The smallest integer satisfying this inequality is the first weak crossover; strict inequality is required for a strict reversal. *Proof.* Lemma 1 substitutes r_k for ρ in both policy values. The phase theorem locates their crossing at $r_k = b$, and r_k increases toward ρ_{op}. When $\rho_{op} > b$, rearranging $\rho_{op}k\rho_a \geq b(1-\rho_a+k\rho_a)$ gives the stated condition. Otherwise $r_k < b$ for every finite k. □This comparison keeps proposal counts fixed and increases annotation spending equally in both arms. It is not an allocation optimum at a fixed total cost.

Sampling with and without replacement. The prior is $p_i = u_i^\kappa / \sum_l u_l^\kappa$ and the budget is $K < N$ proposals. *Without replacement* (the no-repeat instruction), K successive draws come from p restricted to constructs not yet proposed; the value is $V_O = \rho \sum_{i \in S_O} u_i^2$. *With replacement* (restarts), $m \sim \mathrm{Multinomial}(K, p)$ and $V_W = \sum_i u_i^2 \mathrm{SB}_{m_i}(\rho)$. Write $\Delta = \mathbb{E}V_W - \mathbb{E}V_O$.

Lemma 3 (Coupling)**.** *Let $Z_1, Z_2, \ldots$ be i.i.d. draws from p. Let S_W be the n_K distinct values among $Z_1, \ldots, Z_K$, with multiplicities m_i, and let S_O be the first K distinct values of the sequence. Then S_W and S_O have the with- and without-replacement laws, $S_W \subseteq S_O$, $|S_O \setminus S_W| = K - n_K = \sum_{i \in S_W}(m_i - 1)$, and pathwise*

$$V_W - V_O = \underbrace{\sum_{i \in S_W} u_i^2 \{\mathrm{SB}_{m_i}(\rho) - \rho\}}_{\textit{repetition gain}} - \underbrace{\rho \sum_{i \in S_O \setminus S_W} u_i^2}_{\textit{coverage loss}}.$$

Proof. Given the set A of values seen so far, the next new value in an i.i.d. sequence has law p restricted to A^c, which is the successive-sampling transition. The first K draws contain exactly the first $n_K \leq K$ distinct values. The displayed identity then follows by substituting into V_W and V_O. □

Proof of Proposition 3. *(a)* If $\rho = 1$, then $\mathrm{SB}_m(1) = 1$, the repetition gain vanishes and $V_W \leq V_O$ on every path. Now suppose the utilities are exchangeable with respect to the prior; $\kappa = 0$ is one such case. The multiplicity pattern is then independent of the labels, and every distinct slot, in S_W or in $S_O \setminus S_W$, is marginally uniform. Hence $\Delta = \bar{u}_2\{\mathbb{E}\sum_{S_W}[\mathrm{SB}_{m_i}(\rho) - \rho] - \rho\mathbb{E}(K - n_K)\}$, with $\bar{u}_2 = N^{-1}\sum_i u_i^2$. Since $\mathrm{SB}_m(\rho) - \rho = \rho(m-1)(1-\rho)/\{1+(m-1)\rho\} \leq \rho(1-\rho)(m-1)$, it follows that $\Delta \leq -\rho^2\bar{u}_2\,\mathbb{E}(K - n_K) < 0$ for $K \geq 2$. *(b)* Let π_i denote the probability that without-replacement sampling includes construct i. Since the count distribution does not depend on ρ,

$$\frac{\Delta}{\rho} = \sum_i u_i^2 \left\{ \mathbb{E}\left[\frac{m_i}{1+(m_i-1)\rho}\right] - \pi_i \right\},$$

where the term for $m_i = 0$ is zero. Differentiating gives

$$-\sum_i u_i^2 \sum_{m=2}^{K} \Pr(m_i = m) \frac{m(m-1)}{\{1+(m-1)\rho\}^2} < 0.$$

The inequality is strict because $K \geq 2$ and repeated draws of a positive-utility construct have positive probability. Together with the endpoint signs, strict monotonicity gives a unique boundary for $\kappa > 0$; for $\kappa = 0$ the boundary is zero.

Because $\mathrm{SB}_m(\rho) = m\rho + O(\rho^2)$, we have $\lim_{\rho\to 0} \Delta/\rho = K\,\mathbb{E}_p[u^2] - \sum_{r=1}^{K} \mathbb{E}[u_{(r)}^2]$, where $u_{(r)}$ is the utility of the r-th successive draw. The first draw has law p, and by (c) later draws have strictly smaller expected u^2 when $\kappa > 0$, so the limit is positive. As $\kappa \to \infty$ with $u_1 > u_2$, $p \to \delta_1$, so $m_1 \to K$ in probability and the selected utilities converge to the largest K utilities, allowing arbitrary order within ties. Because V is bounded, $\Delta \to u_1^2 \mathrm{SB}_K(\rho) - \rho \sum_{i \leq K} u_i^2$, which is positive if and only if $\rho < \rho_\infty^* = (K/Q_K - 1)/(K-1)$. Two closed forms follow. For a geometric profile $u_i = q^{i-1}$, $Q_K = (1-q^{2K})/(1-q^2)$ and $\rho_\infty^* \to 1 - q^2$ as $K \to \infty$. For $u_i = i^{-\alpha}$ with $\alpha > 1/2$, $\rho_\infty^* \to 1/\zeta(2\alpha)$; with $\alpha \leq 1/2$, $\rho_\infty^* \to 0$, so a flat utility profile leaves no trap in this limit. At $K = 60$, $\rho_\infty^* = 0.20$ for $\alpha = 1/2$ and 0.61 for $\alpha = 1$. *(c)* Condition on the set A of the first $r-1$ draws and let q be p restricted to A^c. The r-th draw has law q. The $(r+1)$-th has law $q_y w_y$, where $w_y = \sum_{x \neq y} q_x/(1-q_x) = C - q_y/(1-q_y)$ decreases in q_y and $\sum_y q_y w_y = 1$. Take any nondecreasing h. Then $h(u_y)$ is nondecreasing and w_y nonincreasing in u_y, because q_y is nondecreasing in u_y. Chebyshev's sum inequality gives $\mathbb{E}_q[hw] \leq \mathbb{E}_q[h]\,\mathbb{E}_q[w] = \mathbb{E}_q[h]$, strictly for strictly increasing h if $\kappa > 0$ and u is non-constant on A^c. Averaging over A completes the proof. With replacement, every draw has law p. *(d)* Under restarts $m_i \sim \mathrm{Binomial}(K, p_i)$, or a sum of independent per-call Bernoulli counts in the realistic design, with p_i increasing in u_i. This family has a monotone likelihood ratio in p_i, so $\mathbb{E}[u \mid m]$ increases in m for any population of constructs, and, for i.i.d. draws, $\log \mathbb{E} m_i = \text{const} + \kappa \log u_i$; in calls of ten it holds only for $p_i \ll 1/10$, and saturation of strong constructs biases a regression estimate of κ downwards. At $\kappa = 0$ the count carries no information about utility. □

Proof of Proposition 1. Write A for the complement of B. Then $\mathbb{E}_q G = b\mu_B + (1-b)\mu_A$ and the constrained law $q(\cdot \mid A)$ has mean μ_A, so the difference is $b\mu_B + (1-b)\mu_A - \mu_A = b(\mu_B - \mu_A)$. Exact duplicates of an observed feature, with no new annotation, add nothing to an oracle predictor, so $\mu_B = 0$ and the difference is $-b\mu_A \leq 0$. The comparison holds per accepted proposal for a fixed history and a fixed proposal law; a natural-language instruction need not implement the conditional law exactly. The difference can be positive even without measurement error: with independent unit-variance factors and $Y = \sqrt{0.45}Z_1 + \sqrt{0.45}Z_2 + \sqrt{0.10}Z_3$, a set holding Z_1 and a family-level ban covering Z_2 trades 0.45 of explained variance for the 0.10 of the more novel Z_3. In the two-layer model of Section 4, a new item of a construct already held has $G = u_i^2\{R(m_i+1,k) - R(m_i,k)\} > 0$ whenever $\rho < 1$ and $u_i > 0$, which is the source of $\mu_B > 0$ analysed by Propositions 2 and 3.

Prediction versus significance counts. Let $X_j = Z + \varepsilon_j$ $(j \leq m)$, where Z, the ε_j and η are independent centred Gaussian variables, with $\mathrm{Var}\, Z = 1$, $\mathrm{Var}\, \varepsilon_j = \sigma^2 > 0$, and $Y = \beta Z + \eta$, $\mathrm{Var}\, \eta = \omega^2$. The best linear predictor has $R_m^2 = \beta^2 m/\{(\sigma^2 + m)(\beta^2 + \omega^2)\}$, increasing in m, while each joint-regression coefficient equals $\beta/(\sigma^2 + m)$. For n independent observations, its squared asymptotic noncentrality is

$$\lambda_m^2 = \frac{n\beta^2\sigma^2}{(\sigma^2 + m - 1)\{\omega^2(\sigma^2 + m) + \beta^2\sigma^2\}},$$

which decreases with m. Additional measurements can therefore improve prediction while reducing power for individual coefficients. A policy can consequently predict better yet produce fewer hypotheses passing a multiplicity-corrected significance test.

Invite-repeat and feedback policies. An agent with memory and permission to repeat chooses, at each proposal, between a new draw and a new item of a known construct. V is separable and concave in the counts, because the marginal gain $u_i^2\{\mathrm{SB}_{m+1}(\rho) - \mathrm{SB}_m(\rho)\} = u_i^2 \rho(1-\rho)/[\{1 + m\rho\}\{1 + (m-1)\rho\}]$ decreases in m. If every construct could be chosen directly with known utility, greedy allocation by marginal gain would be optimal. In our model, however, new constructs arrive through random draws. A draw can also reveal a construct worth measuring again. Re-proposing construct i whenever its immediate gain exceeds the expected gain of a new draw therefore gives a myopic heuristic; it need not be optimal. The invite-repeat arm approximates it with the agent's own importance judgement in place of u. This motivates the hypothesis that invite-repeat can improve on no-repeat in the trap region. Improving on restarts additionally requires useful judgements about

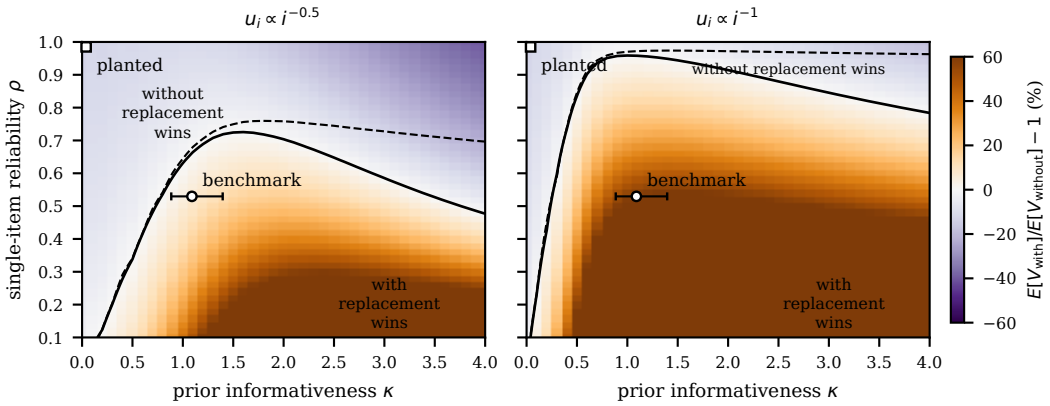

Figure 3: Predicted relative value of proposing with rather than without replacement, for $K = 60$ proposals over $N = 200$ constructs with $u_i \propto i^{-\alpha}$; exact binomial and 40,000 Monte Carlo draws per κ. Solid: zero contour for i.i.d. draws. Dashed: restarts in six calls of ten distinct proposals. Circle: Koijen–Levy estimate with 95% intervals. Square: planted tasks, placed by design at $\kappa = 0$, $\rho \approx 1$ (not estimated).

which constructs to revisit; the model does not prove either comparison. Empirically, invite-repeat recovers 73–74% of the gap and is not significantly below restarts. Score feedback can reweight proposals toward neighbours of high-scoring ideas, represented as $p^{(t)} \propto u^{\kappa} L_t(u)$. Such feedback can be informative or misleading. If informative feedback increases effective κ, then $\partial_\kappa \mathbb{E}_\kappa[u^2] = \mathrm{Cov}_\kappa(u^2, \log u) \geq 0$ and $\mathbb{E}_\kappa[u^2] \leq u_1^2$. Under this representation, expected draw value cannot fall, and the marginal gain tends to zero as $\kappa \to \infty$. Where the marginal gain is largest depends on the utility profile. In the planted tasks of Section 5, feedback supplies information about the unknown mechanism that is absent from the initial proposal policy. In the real-benchmark no-repeat arm, feedback does not eliminate the decline in per-hypothesis utility: the raw-feedback arm still samples without replacement and descends to 0.022. Whether it removes the value loss is unresolved: restarts beat raw feedback in 2026Q1 but not in the sealed quarter (Section 3.2).

Numerics. The phase diagram (Figure 3) uses $N = 200$ constructs, $u_i \propto i^{-\alpha}$ with $\alpha \in \{1/2, 1\}$, and $K = 60$ proposals, the size of one run. $\mathbb{E}V_W$ is exact, because the counts have binomial marginals. $\mathbb{E}V_O = \rho \sum_i u_i^2 \pi_i$ is linear in ρ, and the inclusion probabilities π_i come from 40,000 Gumbel-top-K draws per κ with a fixed seed. The dashed boundary replaces i.i.d. draws by six independent calls of ten distinct proposals each, as in the restart arm. The sign of Δ does not depend on the scale of u. The boundary $\rho_b(\kappa)$ starts at zero at $\kappa = 0$, rises steeply and is hump-shaped: at large κ i.i.d. restarts over-concentrate on the top construct and the solid boundary approaches ρ^*_∞; calls of ten distinct proposals limit the concentration and the dashed boundary stays higher. An adaptive policy could limit this concentration by allocating proposals according to marginal value; invite-repeat is intended to approximate such allocation, but does not guarantee it.

A.1 Calibration on the Koijen–Levy data

This calibration is post hoc. It reuses the three annotations of each Study-2 hypothesis collected for Study 3b, so the model is calibrated on the runs it seeks to explain. In particular, the descent-ratio estimate of κ uses the observed no-repeat decline. The calibration makes no additional model calls. Intervals resample seeds. *Annotation reliability.* For each of 960 items, we average the pairwise correlations among its three 2026Q1 annotations of the identical text. The median is $\rho_a = 0.73$ $[0.72, 0.73]$ (interquartile range across items $[0.62, 0.80]$), consistent with the Study 3a retest 0.70.

Operationalisation reliability. Under the model, two different texts of one construct correlate at $\sqrt{\rho_{a,k}\rho_{op,k}\rho_{a,l}\rho_{op,l}}$, so $\mathrm{corr}(X_k, X_l)/\sqrt{\rho_{a,k}\rho_{a,l}}$ estimates ρ_{op}. We use the 276 cross-round pairs within the near-duplicate clusters of the restart runs (Section 3.2). Their correlation is measured on 2026Q1, not on the 2025Q4 ratings that define the clusters, and is averaged over annotation pairs so that it shares no annotation error. The median between-text correlation is 0.59, and $\rho_{op} = 0.73$

Table 2: Post hoc model sensitivity at $\rho = 0.52965$, $N = 200$ and $K = 60$. Each row uses 60,000 exponential-race draws for inclusion probabilities and exact binomial expectations for repeat counts. Gain is the with-replacement advantage as a percentage of the without-replacement value. "Block" uses six independent calls of ten distinct proposals. These are model calculations, not additional benchmark results.

		Independent draws		Block calls	
α	κ	ρ_b	Gain (%)	ρ_b	Gain (%)
0.5	0.712	0.487	-1.37	0.486	-1.21
0.5	0.804	0.539	$+0.33$	0.541	$+0.34$
0.5	1.090	0.661	$+6.55$	0.671	$+6.07$
1.0	0.712	0.939	$+39.66$	0.946	$+34.98$
1.0	0.804	0.952	$+46.88$	0.960	$+41.37$
1.0	1.090	0.958	$+59.32$	0.972	$+52.32$

$[0.72, 0.74]$ (interquartile range $[0.66, 0.81]$; ratio of medians 0.81; 0.88 on the 11 pairs defined by TF–IDF text similarity alone). Hence $\rho = 0.53$ $[0.52, 0.54]$.

Prior informativeness. First, a zero-truncated Poisson regression of cluster size on the log of the cluster's disattenuated 2026Q1 utility (227 clusters, 77% of them singletons), SIMEX-corrected for sampling error in utility, gives $\hat{\kappa} = 1.1$ $[0.9, 1.4]$ (uncorrected 0.9; 1.1–1.2 for utility floors from 0.005 to 0.02). Second, we match simulated moments under $u_i \propto i^{-1}$. The no-repeat descent ratio 0.34 $[0.24, 0.45]$ (first round 0.050, later no-repeat rounds 0.017, later restart rounds 0.047) gives $\kappa = 0.8$ $[0.7, 1.0]$. The recurrence–utility Spearman 0.51 gives $\kappa = 0.71$ $[0.69, 0.74]$; this interval ignores profile uncertainty. Under $u_i \propto i^{-1/2}$ no $\kappa \leq 4$ reproduces the observed descent (the minimum ratio is 0.39 at finite κ), although the upper end of its interval is attainable; the point estimate therefore favours the steeper profile.

Position in the model. At the direct estimate $\hat{\kappa}$, the original phase-grid calculation gives $\rho_b = 0.66$ ($\alpha = 1/2$) or 0.96 ($\alpha = 1$), both above $\hat{\rho} = 0.53$. Its corresponding restart advantages are 7% and 62% of the no-repeat value (54% in calls of ten), versus 23% observed. These are illustrative model calculations, not a quantitative fit. A post hoc independent reimplementation evaluates the specified parameter values directly, rather than at the nearest phase-grid point (Table 2). Under the steeper profile, all three displayed κ values favour repetition. Under the shallower profile, applying $\kappa = 0.712$ gives the opposite sign and $\kappa = 0.804$ lies close to the boundary. The latter two values were estimated under the steeper profile; applying them to the shallower profile is a sensitivity analysis, not a second moment-matched fit. The phase placement is therefore sensitive to the joint choice of prior informativeness and utility profile.

The annotation crossover is also profile-dependent. At the direct estimate $\kappa = 1.090$, the shallower profile gives $k^* \simeq 3.66$ for independent draws, so four annotations per item cross the boundary; the block calculation gives a first crossing at five. For the steeper profile the boundary exceeds ρ_{op} and no finite annotation count crosses it. These are untested predictions for an intervention that increases annotation spending in both arms.

Model checks. *Study 3a.* By Lemma 1, three annotations of one text raise measured utility by $\sqrt{\mathrm{SB}_3(\rho_a)/\rho_a}$, independently of ρ_{op}. The observed 12.9% against the predicted 11.7% therefore supports the annotation layer but cannot separate the two layers.

Cluster averaging. Lemma 2 predicts that fitting every item equals fitting the cluster means: 3.38 and 3.40 points are observed. It also predicts that averaging each cluster beats keeping one item per cluster by the factor $\sum_c u_c^2 \mathrm{SB}_{m_c}(\rho)/(\rho \sum_c u_c^2)$. The observed factor is 1.11 (3.40 against 3.09), below the two-layer prediction 1.22 and close to the one-layer 1.10. This comparison does not favour the two-layer prediction. One possible explanation is the between-cluster redundancy excluded by (A1): other clusters in the fitted ridge can capture some of the information added by averaging. This explanation is not separately identified.

Decomposition. The restart advantage splits empirically into a distinct-construct part of 0.33 points (restart minus no-repeat, one item per cluster) and a repetition part of 0.31 points (cluster means minus one item per cluster). The additive model predicts the first part to be negative, because

restarts hold fewer distinct constructs (Lemma 3), and it predicts 2.2 points of advantage against 0.63 observed; it reproduces neither the decomposition nor the level. Violations of (A1) and (A5) are possible explanations, discussed in (i) and (v) below. The empirical repetition component accounts for about half the observed advantage. Interpreting that component as re-operationalisation requires the two-layer assumptions; under those assumptions, a re-annotation has 0.51 of the marginal value of a re-proposal.

Study 3b. Re-annotating every item three times scales the distinct-construct part by $\mathrm{SB}_3(\rho_a)/\rho_a = 1.22$ and the repetition part by 0.81 (two layers) or 0.43 (one layer, $\rho_{op} = 1$). With the empirical split, the predicted difference-in-differences is +0.01 points (two layers) or -0.10 (one layer), against the observed 0.00 $[-0.29, 0.30]$. If the whole advantage ran through repetition, the predictions would be -0.12 and -0.36. The point prediction of the one-layer, all-repetition account lies outside the observed 95% interval. This comparison does not propagate uncertainty in the calibrated model parameters and is not a formal model-rejection test. Once the empirical coverage component is included, Study 3b does not distinguish the one-layer and two-layer accounts.

Assumptions and failure modes. (i) *Correlated constructs.* The data depart substantially from (A1). In restart runs the noise-corrected sum of squared utilities is 6.8 times the value of the hypothesis set, and a later hypothesis correlates on average 0.45 with its closest earlier one. The model therefore does not reliably predict levels. Its signs, ratios and orderings should also be interpreted conditional on the stated assumptions. (ii) *Correlated errors across paraphrases* (A3). If the annotator reads paraphrases through a shared bias, the between-text correlation overstates $\rho_a\rho_{op}$. ρ_{op} is then overestimated, and the ceiling on re-proposal falls below one. Error from a shared annotation batch is counted in ρ_{op}. Rating context matters (regrouping events lowers the retest from 0.70 to 0.49), so ρ_{op} need not measure wording effects alone, and a re-annotation in new batches, which Study 3b did not test, could partly substitute for re-proposal. (iii) *Construct definition and conditioning.* Clusters come from a rating-correlation threshold, so $\hat{\rho}_{op}$ is conditional on that threshold. The text-defined estimate is higher but rests on few pairs. Reliabilities use raw rating correlations, whereas utilities condition on the earnings surprise. Applying those reliabilities to the conditional model is an approximation. (iv) *Valid item-specific variance* (A4). If a paraphrase captures a different valid facet, a re-proposal is partly exploration, and the model understates its value. (v) *Finite samples* (A5). Each hypothesis set is fitted by ridge on one quarter of events with one feature per hypothesis. Weak features cost estimation variance, which penalises the no-repeat arm's many weak constructs beyond what the model captures, and selecting among many noisy hypotheses adds a winner's curse. (vi) *Prior form.* $p \propto u^{\kappa}$ indexes how well what the proposer finds salient aligns with what is true. This is a modelling assumption about proposal probabilities, not a claim about a language model's internal sampling process. (vii) *Budget accounting.* All comparisons fix the number of proposals and annotation passes; a re-proposal and a new hypothesis cost the same. *Scope.* The trap requires an informative prior ($\kappa > 0$) and $\rho < \rho_b(\kappa)$, a boundary that depends on K and the utility profile and is zero at $\kappa = 0$; it disappears as $\kappa \to 0$ or $\rho \to 1$. Operationalisation error ($\rho_{op} < 1$) is not required for the trap. It limits how fully re-annotation can substitute for independent operationalisations (Proposition 2). Re-annotation can nevertheless remove the restart advantage if the limiting item reliability exceeds the phase boundary.

B Real-benchmark studies: protocols, prompts, deviations and additional results

Protocol chronology. All real-benchmark experiments were conducted retrospectively on 26 September 2026. The chronology below records when the authors fixed each protocol relative to generating hypotheses, annotating facts and evaluating their outcomes; it does not indicate prospective registration before the market events.

Study 1's protocol specified the arms, splits, models, utility, four endpoint families and inference rule. Its text and code hashes were recorded at 01:16:33 UTC, 0.5–2.5 minutes after the three shared first-round proposals had been generated and rated on 2025Q4. The last protocol edit followed seed 0's first-round ratings. This record preceded all second-round proposals, all 2026Q1 annotations and any evaluation of generated hypotheses on 2026Q1 outcomes. Amendment 1, recorded before unblinding, fixed the official scoring convention (return and surprise ranked within quarter, predictions entered as-is, OLS re-estimated on the evaluation quarter) and added the pooled ridge model and

the increments over the benchmark's baselines. Amendment 2 added the direct-forecast comparator after the primary analysis but before any output from that comparator was available. Amendment 3 added the set-aware arm after the primary analysis; it predicted that this arm would match or beat the no-repeat arm, which it did not.

Study 2's protocol was frozen before its runs. Its analysis and 2026Q1-scoring scripts were hashed about two minutes later, after the first round had been generated and rated on 2025Q4 but before any 2026Q1 annotation. Study 2's restart-minus-no-repeat test (H3 below) was specified as two-sided without a predicted direction.

Study 3a's protocol and analysis script were frozen before its runs; its prespecified prediction that hypothesis-set value rises monotonically with the number of annotations failed at the last step (two annotations score slightly above three; `results_c.json`); it re-annotates hypotheses whose single-annotation 2026Q1 utilities were already known, and its protocol labels it exploratory in design. The novelty-threshold model and the diagnostics of Section 3.2 not listed here were developed after Study 2 and are exploratory.

Study 3b (D1–D3; protocol frozen at 03:05:53 UTC, before its runs) and Study 4 (endpoints 4.1–4.4; 4.5a–4.5b added by an amendment before any 2026Q2 outcome was unblinded to the study analysis) are prespecified. All other analyses are exploratory. 2026Q1 is the evaluation quarter for Studies 1–3. Study 4 evaluated every arm once on 2026Q2, whose outcomes were withheld from the study analysis until 04:17:15 UTC. Its protocol was frozen at 03:36:54, its facts table was built at 03:37:13, and its amendment was recorded at 03:45:00.

Deviations. The following deviations are retained in the study record. (i) Study 1's protocol says the annotator receives one event per call; the frozen code used throughout the studies rates four events per call. (ii) The 2026Q1 annotation driver was edited after annotation had started to shard batches across processes (prompts, model, effort and batch composition unchanged; no outcome had been computed). (iii) Agent, annotator and client code were later extended for the set-aware arm and Studies 2 and 3a; retry limits and the number of concurrent ledger reservations were changed to reduce lock contention (no effect on prompts or models).

(iv) Several annotation processes stopped on transport errors and were restarted; completed batches were kept and the rest re-requested. (v) The self-imposed budget cap of Study 1 was raised from 60 to 80 USD while the set-aware arm was being annotated. (vi) Study 3a's first partial attempt was discarded and the study restarted three minutes after its freeze with the annotator's retry limit reduced to two (prompts, model, effort and batches unchanged).

(vii) Study 4's threshold-policy arm, frozen before 2026Q2 was opened, ran on 2025Q4 in the six minutes after the 2026Q2 facts table was built rather than before it; the 2026Q2 scoring driver was edited after the freeze (scheduling only), the scoring and analysis scripts were re-hashed with the Study-4 amendment, and the invite-repeat scoring driver was hashed ten minutes after the amendment, all before any 2026Q2 outcome was unblinded to the study analysis. The supplement records protocol and script hashes and provides the final scripts. A recorded hash does not establish that every intermediate frozen version is included or can be reconstructed.

Samples. Events come from the benchmark's archived 2025Q4 and 2026Q1 releases, keeping events with facts, an announcement return and an earnings surprise; the sealed 2026Q2 release (Study 4) has 2,060 archived events, of which 1,996 are kept, on 59 label dates. Label dates define clusters. In 2025Q4, every fourth label date (7 dates, 464 events) forms the selection holdout; the remaining 20 dates (1,380 events) are split into an early and a late half by cumulative event count (used by the split arm and the replication selection rule). 2026Q1 has 2,307 events on 59 label dates, 2,199 of them with both archived baseline predictions. The fact summaries carry archive-build timestamps after the events and do not identify the summarising model. An initial keyword screen removed sentences mentioning share prices or market reactions; the inspected removals described pre-event prices, such as buyback prices. A later review identified one retained 2026Q2 summary containing an intraday price reaction. The main text reports a deletion sensitivity for that event. The initial screen therefore missed a case, and the deletion check does not establish point-in-time provenance for the remaining summaries.

Table 3: Explanatory power in the next quarter (2026Q1) and the sealed quarter (2026Q2): official convention, OLS re-estimated on each test quarter; increments over the surprise in points of R^2 with date-clustered t. The pooled model is a ridge, fitted on 2025Q4 only, of the 86 hypotheses with search-set $t > 2$ (for 2026Q2 the same hypotheses, re-annotated in fixed batches on both quarters). Archive baselines are the benchmark's two zero-shot predictions. Direct forecast: GPT-5 (pre-cutoff) asked directly for the return percentile. †Q1: added after protocol recording comparator; Q2: secondary.

	2026Q1		2026Q2 (sealed)	
Predictor	ΔR^2	t	ΔR^2	t
Surprise only (R^2 level)	5.78		7.03	
Archive baseline: GPT-5-nano prediction	5.83	11.2	4.74	12.9
Archive baseline: Gemini Flash-Lite prediction	5.82	9.6	5.79	10.5
Direct forecast† (GPT-5, pre-cutoff)	4.71	9.9	3.53	8.5
TF-IDF ridge on facts (fitted on 2025Q4)	1.67	5.6	–	
Pooled mechanisms	3.96	11.1	3.67	7.8
all coefficients fixed from 2025Q4	3.79		–	
beyond both archive baselines	0.74	4.0	0.20	1.9
beyond the direct forecast†	0.83	4.9	1.00	4.0
beyond both archive baselines and the direct forecast†	0.44	3.4	0.18	1.7

Prompts and arm wording. The proposer receives three inputs. First, a task description specifies the goal (conditions visible in call facts that explain abnormal returns beyond the surprise), how a hypothesis is tested (the annotator's $-2..+2$ rating, then partial correlation with the return rank given the surprise rank), what makes a condition testable, and a non-exhaustive list of useful kinds of mechanisms: forward-looking information (guidance, backlog, bookings, pipeline), earnings quality and persistence (one-offs, margins, cash flow versus accruals), revenue versus cost composition, capital allocation, risk and uncertainty disclosure, expectations management, sector or macro exposure, and interactions. Second, it receives facts and return and surprise percentiles for 30 randomly sampled search events. Third, it receives the arm-specific feedback block. It returns ten hypotheses in a strict JSON schema, each with a condition, sign and mechanism.

The feedback blocks differ as follows. *no-repeat*: earlier conditions, "already tested; do not repeat or trivially rephrase them"; *raw*: each earlier condition with its search-set signed partial correlation, t and incremental R^2, "sorted best first; higher is better"; *split*: early/late half estimates, verdicts and a shrunken estimate, with "build on and refine replicated mechanisms; do not chase noise"; *set-aware*: standalone and incremental utility beyond a ridge of all other hypotheses and the hypothesis set's cross-validated correlation, with "propose mechanisms that add new information"; *restart* (Study 2): no block.

Every arm's closing request asks for ten new hypotheses that are not duplicates or trivial rephrasings of those listed (restart lists none); the invite-repeat arm instead reads that it may re-propose earlier hypotheses in the same or different words. Within a run, conditions with identical text (ignoring case) are annotated once, so invite-repeat runs keep 56–59 of their 60 proposals while every other run keeps 60; the invite-repeat arm therefore retains fewer distinct textual items.

The annotator receives up to ten numbered conditions and four events' facts per call, is told to judge only from the listed facts and not to use outside knowledge of the company or its stock, and returns one integer per condition and event. Conditions are scored on 2026Q1 in exactly the same batches as on 2025Q4.

Pre-specified endpoints not shown in the main text. Study 1 proposal quality (mean next-quarter utility of rounds 2–6): no-repeat 0.0231, raw 0.0273, split 0.0237; raw minus no-repeat 0.0042 $[0.0001, 0.0082]$ (conditional on the generated hypotheses; per-seed differences have one negative seed); split minus raw -0.0036 $[-0.0076, 0.0007]$; split minus no-repeat 0.0006 $[-0.0028, 0.0041]$.

The shared first round averages 0.045; the only later round to match it is raw-feedback round 2 (0.046).

For the top-ten combinations, each run contributes its ten highest-scoring search-set hypotheses, fitted by OLS on 2025Q4 and averaged over seeds: in-sample increments 5.59, 5.56 and 5.30 points; out-of-time 2.61, 2.29 and 1.97.

Single-winner selection does not distinguish the policies: the round-one guidance-raise hypothesis is the search-set best in every run, so search-set, holdout and pooled top-five selection coincide; the cross-fitted next-quarter oracle is also noisy, with a within-2026Q1 odd/even slope of 0.50, so it does not provide a reliable regret estimate. The prespecified holistic annotator rating adds 0.05 points beyond the archive baselines.

Exploratory Study 1 hypothesis-set comparisons. Pooling each arm's three runs (ridge penalty 1000): no-repeat 3.88, raw 3.04, split 3.11 and set-aware 3.39 points over the surprise; 0.78, 0.58, 0.61 and 0.75 beyond the archive baselines. Per seed: no-repeat 3.03, 3.10, 2.75; raw 2.50, 2.66, 2.61; set-aware 2.95, 2.53, 2.46. No-repeat minus raw $0.84\ [-0.11, 1.75]$; set-aware minus no-repeat $-0.48\ [-1.62, 0.39]$, minus raw $0.35\ [-0.45, 1.00]$.

The no-repeat-over-raw hypothesis-set advantage did not replicate in Study 2 (H1: -0.12 points, 2 of 6 seeds positive, $p = 0.793$).

Study 2 details. Seeds 3–8; each arm makes six proposer calls and keeps up to 60 hypotheses; all arms share the first call and the seed×round inspection samples. Prespecified one-sided tests: H1 (no-repeat minus raw, hypothesis set) -0.12, $p = 0.793$; H2 (raw minus no-repeat, per hypothesis) 0.0047, 5 of 6 positive, $p = 0.022$; two-sided H3 (restart minus no-repeat, hypothesis set) 0.63, 6 of 6, $p = 0.007$.

Arm-pooled hypothesis sets beyond the archive baselines: restart 0.93, no-repeat 0.62, raw 0.70 points. Nine-seed (Studies 1 and 2) H1: 0.04, $p = 0.365$; H2: 0.0046, $p = 0.007$. Prespecified secondary test (date-cluster bootstrap of arm-pooled differences, 1,000 draws): restart minus no-repeat $0.81\ [-0.12, 1.66]$, restart minus raw $0.67\ [-0.06, 1.46]$, no-repeat minus raw $-0.14\ [-0.73, 0.58]$. The restart-minus-raw seed contrast and restart's per-hypothesis advantage were not prespecified.

Study 3b (measurement reliability). The protocol was frozen at 03:05:53 UTC before any runs. The USD 200 budget was approved at 03:09:44, followed by the first annotation call at 03:09:52. Each Study-2 hypothesis received two additional annotations on 2025Q4 and 2026Q1, with batches held fixed. Hypothesis-set values use the mean of k annotations. Prespecified D1 (one-sided, predicted negative): the difference-in-differences of the restart advantage between $k = 3$ and $k = 1$ is 0.00 points over no-repeat ($p = 0.509$; 3 of 6 seeds negative) and -0.03 over raw ($p = 0.361$), so the prediction failed.

D2: with three annotations restarts still beat no-repeat by 0.64 points (6 of 6 seeds, $p = 0.003$) and raw by 0.48 (6 of 6). D3: mean per-hypothesis utility rises by factors 1.10 (restart), 1.22 (no-repeat) and 1.09 (raw). Mean hypothesis-set values with one and three annotations: restart 3.38 and 3.67, no-repeat 2.75 and 3.03, raw 2.86 and 3.19. Part of the repeat annotation ran on a second API account after the first account's ledger reached its authorised cap; completed files were kept.

Study 4 (sealed-quarter confirmation). Its prespecified endpoints are 4.1, the pooled-mechanism increments (Table 3); 4.2, restart minus no-repeat and restart minus raw; 4.3, raw minus no-repeat per hypothesis; 4.4, threshold minus restart; and, added by amendment, 4.5a, invite-repeat minus no-repeat, and 4.5b, restart minus invite-repeat (Table 1). The protocol, the analysis script, the threshold-policy arm and its code were frozen before the 2026Q2 archive was read. The facts table was built first and annotated; the same step wrote returns, surprises and archived predictions to a separate sealed file, retaining events with non-missing return and surprise without printing outcome values. The preparation step therefore parsed these outcomes; their values were first unblinded to the study analysis after all annotation had finished, at 04:17:15 UTC. The Study-4 amendment, recorded after the facts had been opened for annotation but before any outcome was unblinded to the study analysis, added the invite-repeat arm and its two two-sided tests. Endpoint 4.1 uses Study 3a's first repeat annotation of the 86 selected hypotheses (fixed batches) on 2025Q4 for fitting and the same batches on 2026Q2. The direct forecast failed on 11 of 1,996 events, set to the median as specified in the protocol.

The threshold policy explored in rounds 2–3 of every seed and re-annotated in 30% of rounds 2–6; its per-seed sequences are in the supplement. Per-seed hypothesis-set means in 2026Q2: restart 3.52, no-repeat 2.95, raw 3.36, invite-repeat 3.37, threshold 3.27 points. Prespecified endpoint 4.3 (raw minus no-repeat, per hypothesis, nine seeds): 0.0031 ($p = 0.107$). Across Study-1 hypotheses, 2026Q2 utility regresses on search-set utility with slope 0.48 and correlates 0.58 with 2026Q1 utility.

Novelty-threshold check. With $\rho = 0.70$ (identical batches) the threshold is $\tau = 0.42$; with $\rho = 0.49$ (regrouped events) it is 0.59. Pooling Studies 1 and 2, later-round proposals average, relative to their run's best first-round hypothesis, 0.14 (no-repeat), 0.17 (raw), 0.18 (split) and 0.34 (restart), and 10%, 15%, 16% and 37% of them clear $\tau = 0.42$. Study 3a: mean per-condition next-quarter utility 0.048 with one annotation and 0.055 with three (Spearman–Brown prediction 11.7%, observed 12.9%); pooled increments 3.95, 4.19 and 4.16 points for one, two and three annotations.

Table 4: Ten highest-ranked hypotheses meeting both screening criteria: search-set $t > 2$ and next-quarter FDR 10%, ranked by in-sample (2025Q4 search-set) utility; inclusion requires next-quarter significance, so the 2026Q1 column is conditioned on the evaluation quarter. Near-duplicates (rating correlation above 0.5 or similar wording) are dropped. Conditions are the agents' wording, lightly truncated.

Sign	Condition	Q4	Q1
+	Company raises current-year guidance (revenue, EPS, or EBITDA) during the call.	0.180	0.149
+	Management highlights unusually high free cash flow conversion (e.g., ≥80% of adjusted EBITDA) alongside low…	0.120	0.103
−	Revenue mix shifted toward explicitly lower-margin categories, with gross margin down year-over-year due to mix.	0.116	0.060
−	Management cited elevated customer/channel inventories that are depressing sell-in and will take time to normalize.	0.114	0.087
−	Management attributed material revenue pressure to third-party platform policy changes or identified…	0.107	0.084
+	Management reports trailing-12-month free cash flow turning positive for the first time and guides to continued…	0.105	0.057
+	Leverage and liquidity improved (e.g., net debt reduced, voluntary debt repayments, cash balances increased),…	0.103	0.065
−	Management withdrew or declined to provide near-term guidance due to uncertainty or external disruptions.	0.099	0.099
−	Management cites unfavorable fixed-cost absorption leading to gross margin pressure or negative gross margins.	0.096	0.085
+	The company reported positive operating leverage (revenue growth outpacing expense growth) or explicitly cited…	0.096	0.111

Guidance and non-guidance content. Of the 86 selected hypotheses, 23 mention guidance, outlooks, forecasts or targets (keyword rule); pooled, they add 3.16 points over the surprise, and the other 63 add 2.90 (0.50 beyond both archive baselines).

Cost. Settled usage-based cost: Study 1 (proposer, annotator on both quarters, holistic rating and the set-aware arm) USD 54.60; the direct forecast USD 9.79; Study 2 USD 81.67; Study 3a USD 23.39; Study 3b USD 149.29; Study 4 (threshold and invite-repeat arms, all 2026Q2 annotation and the direct forecast) USD 150.93; the retest-reliability check USD 0.16; a smoke test USD 0.01. 226 reservations from interrupted processes (upper bound USD 47.68) were never settled and remain conservatively reserved in the ledger.

B.1 Benchmark leaderboard studies (Study 5)

Split, models and timeline. Training events are the 2025Q4 announcements dated on or before 4 November 2025 (1,028 events, 19 label dates); the evaluation sample is the benchmark's late half (816 events on 9 dates: the archive split at 4 November gives 1,031 early and 818 late events, against 1,032 and 817 reported, one boundary event differing; dropping the five events without facts leaves 1,028 and 816). Model knowledge cutoffs (provider documentation): GPT-5.2 2025-08-31; GPT-5 and GPT-5.1 2024-09-30; o3, o4-mini and GPT-4.1 2024-06-01; the annotator GPT-5-mini 2024-05-31; all precede the first event (2025-10-09).

The holistic predictor's model and effort (GPT-5.2, low reasoning, one sample) were chosen on 400 training events, where test–retest correlation was 0.96–0.97. Study 5a's protocol and code were hashed at 04:49:35 UTC, its ACE-lite schedule was amended from 6×40 to 8×30 training examples

(to respect the per-call budget cap) before evaluating any outcome, and its analysis script was hashed at 04:53:19; its outcomes were computed at about 05:16.

Study 5b's protocol and analysis were hashed at 05:25:56 while its predictions were running and before any of its outcomes; during the runs the predictor script gained a resume-from-receipts option and accepted a bare-integer reply returned by some o3 calls, and the three GPT-5 runs were restarted with more concurrency (no prediction changed). Study 5b's outcomes were computed at 06:07.

Study 5c's protocol and analysis were hashed at 06:09:39, before any of its predictions; its outcomes were computed at 06:27. Studies 5b and 5c were designed after Study 5a's outcomes and are post hoc; the 2026Q1 and 2026Q2 outcomes had also been read in earlier studies.

Playbooks. ACE-lite: eight rounds; in each, GPT-5.2 (medium reasoning) sees the current playbook and 30 training events with their return and surprise percentiles, reflects on which patterns explain returns beyond the surprise, and returns an updated playbook of at most 25 bullets (final: 24 bullets).

Tested-mechanism playbook: 24 memoryless restarts of the Section 3.1 proposer (GPT-5.2, medium; inspection samples from training events only) gave 240 hypotheses; 136 had training $t > 2$ in their stated direction, 103 remained after merging near-duplicates (rating correlation above 0.7), and the 25 with the largest t form the playbook, each stated as a condition, a direction, an evidence grade and a mechanism.

Under the keyword rule *guid, outlook, forecast, next quarter, upcoming*, 9 of the 24 ACE-lite bullets and 16 of the 25 tested bullets concern forward guidance or outlook; the counts depend on the rule (a broader rule that also counts near-term and visibility language classifies most bullets of both playbooks as forward-looking). M is a ridge on the 103 kept ratings (penalty chosen by date-grouped cross-validation on training events). Both playbooks and all prompts are in the supplement.

Study 5c prompts. Conditional framing (primary): the predictor is told the announcement's surprise percentile against the training distribution and asked for "the percentile of this stock's return among companies with a similar surprise (50 = the typical return for this surprise)". Total framing: the same surprise line, asking for the return percentile among all companies; its predictions track the surprise (correlation with the surprise rank 0.67–0.89 on training events) and score lower than the conditional framing.

Individual predictions. Table 5 reports each of the 18 single instruments of Study 5b; the composites and increments with intervals are in the supplement (`results_f.json`, `results_g.json` and `results_h.json` for Studies 5a, 5b and 5c). Settled cost of Study 5: USD 524.

Table 5: Joint R^2 (%) of each single instrument (Study 5b): model × context, on the benchmark's late sample, 2026Q1 and 2026Q2.

	Zero-shot			ACE-lite playbook			Tested playbook		
Model	Late	Q1	Q2	Late	Q1	Q2	Late	Q1	Q2
gpt-5.2	11.4	10.6	11.5	12.0	11.5	12.1	11.8	11.1	11.5
gpt-5.1	10.4	9.9	10.0	10.5	9.8	10.8	10.6	10.3	10.7
gpt-5	10.0	10.1	10.3	10.9	10.7	10.5	10.8	11.1	11.1
o3	11.1	10.7	10.8	12.3	10.9	10.6	11.7	10.7	10.6
o4-mini	10.7	9.7	10.6	10.9	9.5	11.0	11.3	9.7	11.2
gpt-4.1	10.8	9.3	10.7	10.9	10.0	11.1	10.9	10.0	11.1
M (mechanism score)	11.2 / 9.7 / 10.7								

B.2 Joint run and date uncertainty

This review-stage analysis was specified after the primary results were known. It supplements the original paired-run tests by varying both generated-run blocks and evaluation dates. For each quarter, 5,000 bootstrap draws independently resample the six paired runs and 59 date clusters with replace-

ment. Within a draw, every arm uses the same sampled runs and dates. The original hypotheses, 2025Q4 ridge weights and quarter-level outcome and surprise ranks remain fixed. The test-quarter calibration OLS is re-estimated on each sampled set of dates before computing the mean paired-run contrast. Thus this analysis quantifies uncertainty conditional on the generated hypotheses and training data; it is not a fresh-run replication or a fully frozen-head evaluation.

Table 6: Review-stage joint run/date bootstrap: unadjusted 95% percentile intervals for mean differences in incremental R^2, in percentage points. Both comparisons and quarters are shown.

Comparison	2026Q1	2026Q2
Restart minus no-repeat	$[0.123, 1.195]$	$[0.057, 1.067]$
Invite-repeat minus no-repeat	$[-0.071, 1.044]$	$[-0.104, 0.836]$

The restart comparison remains positive under these intervals in both quarters. The invite-repeat intervals include zero, although its paired-run point estimates are positive. The two results therefore warrant different uncertainty statements: the invitation result demonstrates improvement across the observed runs, but remains less precise when evaluation dates are also resampled. These exploratory intervals are unadjusted for multiple comparisons and do not replace the originally specified tests. The supplement provides a portable replay script and aggregate outputs.

C LEADERBOARD COMPARISONS (STUDY 5)

Table 7: Leaderboard comparisons under the metric of Koijen & Levy (2026): joint R^2 (%) from regressing return ranks on surprise ranks and a prediction, with coefficients re-estimated on each evaluation sample. Late 2025Q4 contains 816 announcements on 9 dates; fitting uses the earlier half. Published rows are reproduced as reported, including systems using Opus 4.5; our experiments use only OpenAI models. Intervals use a date-cluster bootstrap. Our prespecified analysis ranks returns and surprises within each evaluation sample. Published rows instead use whole-quarter ranks, and their surprise-only value refers to the full quarter. Rows using archived predictions retain events with both predictions: 2,199 in 2026Q1 and 1,904 in 2026Q2. Studies 5b and 5c were designed after Study 5a; their code was hashed before their respective outcomes were computed.

System	Late 2025Q4 (benchmark sample) R^2	95% interval	2026Q1 R^2	2026Q2 R^2
Reported by Koijen and Levy				
Surprise only (full quarter)	5.0		–	–
GPT-5-nano, zero-shot	8.5		–	–
Gemini 2.5 Flash-Lite, zero-shot	10.5		–	–
GEPA (Gemini Flash-Lite)	12.8		–	–
Opus 4.5, zero-shot	14.1		–	–
ACE (Opus 4.5)	17.1		–	–
Study 5a (pre-registered): GPT-5.2				
Zero-shot	11.4	[6.8, 15.2]	10.6	11.5
ACE-lite playbook (best practice)	12.0	[6.5, 16.7]	11.5	12.1
Tested-mechanism playbook (ours)	11.8	[6.8, 16.0]	11.1	11.5
Study 5b (post hoc): six pre-cutoff models, equal-weight average				
ACE-lite playbooks (C_{ACE})	12.0	[6.1, 16.7]	11.1	11.6
Tested playbooks (C_{ours})	11.8	[6.6, 16.2]	11.1	11.5
Tested playbooks + M (G^*, ours)	12.1	[6.7, 16.6]	11.2	11.7
G^* + archived predictions	12.6	[7.1, 16.9]	12.3	12.0
Study 5c (post hoc): GPT-5.2 given the surprise				
G_H: tested playbook + M (ours)	12.6	[6.7, 17.0]	11.1	11.6

Design. We use the benchmark's early/late split in 2025Q4 and fit all predictors on the 1,028 early events. We then evaluate on late 2025Q4, 2026Q1 and 2026Q2 using the metric in Table 7. The

holistic predictor receives event facts and, in some arms, a playbook, and returns a predicted return percentile. All prediction models have provider-reported knowledge cutoffs before the events.

The prespecified *Study 5a* uses GPT-5.2 to compare zero-shot prediction with two playbooks. The ACE-style control ("ACE-lite") curates at most 25 bullets over eight rounds of 30 training events. It omits ACE's generator, reflector and delta updates (Zhang et al., 2026a). The tested-mechanism arm generates 240 hypotheses through 24 memoryless restarts. After filtering at training $t > 2$ and merging near-duplicates, a ridge on the 103 remaining ratings defines the mechanism score M; the 25 hypotheses with the largest t form its playbook.

Study 5b averages training-standardised predictions with equal weights across six models: GPT-5.2, GPT-5.1, GPT-5, o3, o4-mini and GPT-4.1. It forms a separate ensemble for each context. *Study 5c* supplies GPT-5.2 with the earnings surprise and asks for the information in the facts beyond that surprise. Appendix B.1 gives the prompts and protocols.

Playbooks produce similar predictions. GPT-5.2 zero-shot achieves 11.4% on the late sample, exceeding the archived GPT-5-nano and Gemini predictions there but falling below them in 2026Q1–Q2. ACE-lite adds +0.5 points ([−0.6, 1.7]). After removing the surprise, GPT-5.2 predictions under the three contexts correlate 0.90–0.93 on training events. This similarity is consistent with the playbooks emphasizing information that the zero-shot model already uses, as suggested by the informative-prior interpretation in Section 3.2; it does not identify the model's internal reasoning.

The prespecified comparison does not establish that the tested playbook outperforms ACE-lite: their difference on the benchmark sample is −0.2 ([−1.1, 0.8]). The differences are negative in 2026Q1 and 2026Q2 (−0.4 and −0.5; the latter interval is [−1.14, −0.01]). The stacked system assigns zero weight to M. Across six models, both playbooks improve the average instrument from 10.7% to 11.2% for ACE-lite and 11.2% for the tested playbook. Supplying the surprise in Study 5c leaves little difference between them (+0.2 points); a nonlinear surprise map yields 4.8%, without further improvement.

Ensembling and explicit mechanism scores. Averaging six models adds +1.1 points ([0.2, 2.2]) over the single model selected on training events, but does not improve on GPT-5.2 alone (12.0% versus 12.0%). Contrary to the low-correlation expectation in Study 5b's protocol, the six models' surprise-residualised predictions correlate 0.88 on training events. The correlated-error Spearman–Brown calculation (Section 4) uses training data to predict a composite gain factor $\sqrt{J/\{1+(J-1)\bar{r}\}} = 1.05$ for J instruments with mean correlation $\bar{r}$; the realised factors are 1.06–1.08, above the prediction in 11 of 12 composite–sample cells (the remaining cell falls short in the fourth decimal).

The explicit mechanism score is less correlated with the holistic predictions (0.74–0.75, against 0.90–0.93 among them) and adds to the averaged tested playbooks in all three samples (+0.3, +0.2 and +0.1 points), significantly only on the late sample ([0.03, 0.44]).

Limits of leaderboard comparisons. Under the benchmark's quarter-rank convention the archived GPT-5-nano and Gemini predictions reproduce the reported 8.5% and 10.5% (8.5% and 10.5%), and our best system without archived inputs (G_H) scores 12.2% (12.6% with our within-sample ranks); the reported GEPA scores 12.8%, and Opus 4.5 14.1% zero-shot and 17.1% with ACE. The 9 dates do not resolve a ranking of these systems. Even the 1.7-point gap between G_H and archived Gemini predictions on the same events has interval [−0.8, 3.5]. Koijen and Levy report no paired tests for their systems. We therefore base our conclusions on paired increments within a model. In the later quarters our systems score 9.3–12.2%, against 11.6–12.7% for the archived predictions. Testing whether explicit mechanism scores improve Opus 4.5 with full ACE remains an open comparison.

D PLANTED-TRUTH EXPERIMENTS

The planted tasks (Section 5) use point-in-time features on an asset×date panel, labels $y = \mu(X) + \varepsilon$ with a planted mechanism μ from five families (linear, interaction, threshold, event and delayed flow) and noise calibrated to daily returns. A candidate is a typed expression program with one

fitted coefficient. Searchers observe noisy search-set gains and may spend evaluation queries on independent validation scores for shortlisted candidates. Only the evaluator observes noiseless utility on a separate design. The proposal policy has no task-specific information about μ before receiving feedback. Experiment P1 studies feedback, P2 the precision of validation and P3 the transfer of learned selectors.

D.1 P1: FEEDBACK AND DISCOVERY

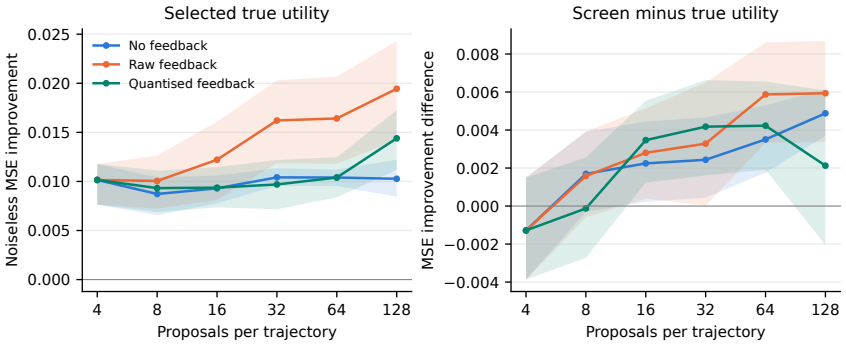

Figure 4: P1 (planted truth): selected true utility (left) and search–truth gap of the selected program (right) against search depth. Task-stratified world-bootstrap bands, conditional on the recorded policy tables.

Design and prespecified contrasts. The complete main design uses the frozen editor at medium effort, with 2 policy tables, 3 tasks, 6 noise worlds and 128 proposals per trajectory. Policy tables are averaged before task-stratified paired bootstrap over worlds. Contrast (a), the raw-channel gap slope per doubling, is $+0.00141$ ($[+0.00079, +0.00198]$); the first checkpoint exceeding twice the early mean gap is undefined (nonpositive initial mean gap).

Contrast (b) gives raw minus no feedback $+0.00917$ ($[+0.00476, +0.01313]$, approximate Holm-adjusted tail probability < 0.001), and Ladder minus raw, where Ladder feedback reveals a search-set score only when it beats the best so far (Blum & Hardt, 2015) -0.00505 ($[-0.01160, +0.00206]$, 0.324).

Contrast (c) gives independent top-k minus search-set selection $+0.00013$ ($[-0.00049, +0.00071]$, 0.656); widening the correction to all feedback channels gives 0.867. Contrast (d), the oracle-minus-selected regret, is $+0.00470$ for no feedback, $+0.00394$ for raw and $+0.00210$ for Ladder.

These estimates condition on the fixed task designs and memoized policy tables. A growing search–truth gap does not by itself imply declining true utility. Technical dry-run observations are excluded. Raw feedback improves the selected program's true utility despite a growing search–truth gap (Figure 4); the fixed holdout rule does not establish an additional benefit. The three-model ordering is unavailable: `gpt-5.4-mini` failed the output contract and part of `gpt-5.6-sol` stopped at its fixed cost reservation cap.

P1 accounting appears in Table 8; the complete batch recorded $45.5844. All terminal failures and full protocol hashes are retained. `gpt-5.4-mini` stopped at its output ceiling; two `gpt-5.6-sol` tasks stopped at the conservative pre-dispatch cost reservation cap. These incomplete grids do not support the planned capability ordering.

D.2 P2: VALIDATION PRECISION AT A FIXED QUERY ALLOWANCE

This follow-up was designed after the original P2 results. Its protocol, source hashes and analysis were frozen before generating the new seed outcomes. The 64 seeds (492300–492363) are disjoint from the seeds found in 17 archived synthetic protocols and the documented seed ranges of the other planted experiments. Each seed has four noise worlds in each of the original four families, including the zero-signal family. Technical tests used a separate seed. The study varies validation-panel size while holding the searcher, grammar, fitted coefficient, search-set panel and shortlist rule fixed.

Table 8: P1 terminal accounting. Incomplete grids are excluded from model comparisons. Protocol digests are SHA-256 prefixes; full digests and all terminal summaries are in the archived mirror.

Model / table	Task	Rounds	Recorded USD	Protocol digest
5.6-luna / 0	linear_0.08	32	0.5347	`7bad18d35d29`
5.6-luna / 0	abs_diff_times_0.25	32	0.7826	`77786b0e5a03`
5.6-luna / 0	max_times_0.25	32	0.7734	`582aa7ef234e`
5.6-luna / 1	linear_0.08	32	0.5703	`308ca96750d7`
5.6-luna / 1	abs_diff_times_0.25	32	0.8312	`901c80390422`
5.6-luna / 1	max_times_0.25	32	0.7637	`f0854a198104`
5.4-mini / 0	linear_0.08	9	0.8127	`91f2481eca9c`
5.4-mini / 0	abs_diff_times_0.25	3	0.2642	`25e229ad3602`
5.4-mini / 0	max_times_0.25	6	0.5907	`7aefd7c0f63d`
5.6-sol / 0	linear_0.08	32	11.2559	`7d580629b3d3`
5.6-sol / 0	abs_diff_times_0.25	30	14.1251	`2de3c9c01c31`
5.6-sol / 0	max_times_0.25	30	14.2800	`2578962d22f3`

Table 9: P2 precision resource accounting. Independent selection reserves 8 of 128 queries. Row evaluations count search-set plus validation; training and evaluator-only truth work are separate. All entries derive from the frozen precision-study macros.

Policy / validation dates	Query allowance	Distinct labels	Row evaluations
All-search	128	768	32768
Independent, 2 dates	128	800	30976
Independent, 8 dates	128	896	31744
Independent, 16 dates	128	1024	32768
Independent, 64 dates	128	1792	38912

A common search trajectory. For each seed/family/world, raw-feedback hill climbing generates one 128-query trajectory. The all-search control uses that trajectory; both independent selection and its same-prefix search-set control use the first 120 queries. The 8 distinct candidates with the best search-set scores form a shortlist before validation scores are computed. Independent selection evaluates that shortlist using $h \in \{2, 8, 16, 64\}$ validation dates; a zero/no-edit alternative remains free. Training, search-set and noiseless truth arrays are identical across settings. Validation scores never affect proposals. The original simulator draws panel features before panel noise, so outcomes at different h are not nested samples. The experiment compares marginal performance under the same data-generating process at different panel sizes, not the effect of appending dates to a fixed observed panel.

Query, information and computation costs. Each policy has a 128-query allowance, but this does not equalize evidence or computation. All policies use 512 training and 256 search-set labels. Independent selection additionally uses $16h$ distinct validation labels, shared by the 8 shortlisted candidates. The four settings therefore use $800, 896, 1024, 1792$ distinct labels, respectively, compared with 768 for all-search. Their charged search-set-plus-validation row evaluations are $30976, 31744, 32768, 38912$, compared with 32768 for all-search.

Training work, reusable program caches and evaluator-only truth computations are recorded separately. The 16 assets on each date share date-level noise; label rows are not independent observations. A gain at larger h cannot be attributed to a better algorithm under a fixed label or compute budget.

Inference and interpretation. The analysis first averages worlds within family, then equally weights the four families within seed. Joint bootstrap resampling of the 64 seeds retains every family and precision setting (20000 resamples). All 40 prespecified family/aggregate contrasts are reported below; the 8 aggregate comparisons additionally use Bonferroni 99.375% marginal percentile intervals. These are approximate bootstrap intervals, not exact finite-sample guarantees. The repeated all-search and same-prefix controls across h are shared observations, not additional replications. The gain always equals the change in pool-oracle utility minus the change in selection regret.

At $h = 16$, the net-gain point estimate has the same sign as the original P2 result, but its ordinary 95% interval crosses zero. The same-prefix selection contrast is positive at 95%, while both wider intervals cross zero. Thus this follow-up does not independently establish the original net benefit. At $h = 2$, both controls outperform independent selection even under the wider intervals; at $h = 8$, the net contrast is negative under the wider interval, while the same-prefix interval narrowly crosses zero. At $h = 64$, both contrasts are positive under the wider intervals. The estimated tradeoff for holdout-only selection changes sign across these settings. The experiment does not locate an exact threshold, establish a universal scaling law, or demonstrate that a research procedure was learned. The post hoc control below changes the selector while holding its purchased evidence fixed.

Table 10: All 8 aggregate contrasts, in 10^{-3} squared-label utility units. Entries are estimates and Bonferroni 99.375% seed-bootstrap intervals.

Dates	Versus all-search	Versus prefix search set
2	-2.7466 [-3.6649, -1.8717]	-2.6518 [-3.5537, -1.8141]
8	-0.5572 [-1.1080, -0.0528]	-0.4624 [-0.9503, +0.0023]
16	+0.2322 [-0.1519, +0.5997]	+0.3270 [-0.0409, +0.6788]
64	+1.1031 [+0.8017, +1.3910]	+1.1979 [+0.9576, +1.4638]

Table 11: Every family-specific precision contrast, including null; estimates and unadjusted 95% seed-bootstrap intervals, in 10^{-3} squared-label utility units. These secondary contrasts are not multiplicity-adjusted.

Dates	Family	Versus all-search	Versus prefix search set
2	Absolute interaction	-7.0436 [-8.8559, -5.3327]	-6.8023 [-8.6024, -5.1105]
2	Linear	-0.7064 [-0.9733, -0.4553]	-0.7210 [-0.9867, -0.4702]
2	Maximum interaction	-3.6925 [-5.0130, -2.4626]	-3.5374 [-4.7930, -2.3704]
2	Null	+0.4561 [+0.3568, +0.5584]	+0.4535 [+0.3533, +0.5575]
8	Absolute interaction	-1.4126 [-2.2737, -0.5923]	-1.1713 [-1.9540, -0.4254]
8	Linear	-0.3509 [-0.6216, -0.0864]	-0.3656 [-0.6328, -0.1002]
8	Maximum interaction	-1.0241 [-1.9132, -0.1593]	-0.8689 [-1.6722, -0.0578]
8	Null	+0.5589 [+0.4590, +0.6637]	+0.5564 [+0.4572, +0.6603]
16	Absolute interaction	-0.1846 [-0.8814, +0.4571]	+0.0567 [-0.6205, +0.6882]
16	Linear	+0.0433 [-0.2075, +0.2779]	+0.0287 [-0.2182, +0.2591]
16	Maximum interaction	+0.4016 [-0.1716, +1.0020]	+0.5567 [+0.0129, +1.1375]
16	Null	+0.6684 [+0.5550, +0.7811]	+0.6658 [+0.5509, +0.7808]
64	Absolute interaction	+1.3080 [+0.8484, +1.7674]	+1.5492 [+1.1586, +1.9690]
64	Linear	+0.6536 [+0.4588, +0.8523]	+0.6390 [+0.4484, +0.8322]
64	Maximum interaction	+1.4795 [+0.9240, +2.0687]	+1.6346 [+1.1407, +2.1775]
64	Null	+0.9712 [+0.8547, +1.0884]	+0.9686 [+0.8553, +1.0843]

POST HOC POOLED-SCORE CONTROL

Table 12: P2 same-evidence control: sample-size-pooled minus holdout-only selected utility (10^{-3}), 64 seed clusters, 98.75% Bonferroni intervals; posthoc.

Validation dates h	Pooled minus holdout-only
2	$+2.786\ [+2.049, +3.583]$
8	$+0.856\ [+0.539, +1.206]$
16	$+0.292\ [+0.067, +0.559]$
64	$-0.079\ [-0.137, -0.019]$

The precision study's final selector ranks shortlisted candidates using their independent validation scores alone. An ordinary alternative retains the adaptively used search-set score. After inspecting the original precision results, we fixed

$$\widehat{U}_{\text{pool}}(f) = \frac{n_s \widehat{U}_s(f) + n_v \widehat{U}_v(f)}{n_s + n_v}, \qquad n_s = 256, \quad n_v = 16h.$$

Both terms are mean squared-error improvements of the same train-fitted prediction over zero, so this is the empirical gain on the combined panels. Because the search-set score was used to build the shortlist, the pooled score need not be an unbiased or optimal estimator of utility. The rule and sample-count weights are ordinary controls, not a proposed methodological innovation.

All 64 existing seeds, the original families and worlds, and every $h \in \{2, 8, 16, 64\}$ are retained. The shortlist is the first occurrence of each distinct program in the 120-query prefix, stably ranked by search-set score; only the top 8 candidates are eligible. The zero predictor is considered first and strict improvements determine selection. No oracle utility enters the weights, shortlist or tie rule. The four precision settings produce 4,096 new pooled selections. All original policy choices and utilities are reproduced exactly.

The primary comparison within this post hoc analysis is pooled minus holdout-only at each panel size (Table 12). Joint resampling of the same seed clusters gives ordinary intervals and 98.75% Bonferroni marginal intervals across its 4 aggregate contrasts. Families and worlds are averaged with the original equal weighting. These intervals describe the historical sample; multiple-comparison adjustment does not remove the post hoc choice of control. No new seeds, candidate generation or model calls are involved.

Pooling outperforms holdout-only selection at the smaller panel sizes and underperforms it at the largest; the stated wider intervals exclude zero for these contrasts. Relative to all-search, its smallest-panel estimate is close to zero with an ordinary interval crossing zero; the other pooled estimates are positive (Table 13). Those secondary contrasts do not jointly match information and evaluation costs, and their intervals are unadjusted. Thus the original holdout-only sign reversal survives as a policy-specific observation, while its small-panel losses are sensitive to a simple use of the already purchased scores.

Neither pooled nor holdout-only selection dominates across panel sizes. Selecting whichever policy won at each observed h would be an oracle choice, not a learned or prospectively validated policy.

Table 13: Post hoc secondary contrasts, in 10^{-3} squared-label units, with ordinary 95% intervals. All-search has the same query allowance but fewer labels; same-prefix search-set does not purchase the validation scores. These are not the same-evidence primary contrasts. Full family-level outcomes are retained in the computational supplement.

h	Pooled minus all-search	Pooled minus same-prefix
2	$+0.040$ $[-0.121, +0.191]$	$+0.135$ $[+0.042, +0.236]$
8	$+0.299$ $[+0.077, +0.517]$	$+0.393$ $[+0.231, +0.569]$
16	$+0.524$ $[+0.330, +0.712]$	$+0.619$ $[+0.466, +0.780]$
64	$+1.024$ $[+0.816, +1.227]$	$+1.119$ $[+0.954, +1.290]$

The execution record includes an initial parser failure involving the literal null-family name and a later repair to summary serialization. The latter occurred after selection and scoring; all six numerical CSV outputs before and after that repair are byte-identical. The executed protocol and original selection seal are preserved separately from later chronology notes. These implementation repairs did not change weights, candidates, choices or numerical outcomes.

D.3 P3: FINITE-EVIDENCE SELECTOR TRANSFER AND REJECTION CONTROLS

This study was designed after the P2 precision and pooling results. Its source models were sealed before the new target instances were generated, and all target choices were sealed before auditor utility was decoded. Integrity checks had read the truth-file bytes to compute hashes. The record therefore establishes a separation in data flow, not operating-system isolation. The post hoc rejection controls described below reuse these now-exposed targets and are not a second fresh confirmation.

Source and target episodes. Source seeds 492300–492363 reuse the original hill-climbing trajectories: the first 32 seeds fit and the remaining 32 tune. Within the first 120 proposals, the first occurrence of each program is retained and the top 8 search-set scores form a stable shortlist. Each candidate has its original training-only coefficient and centering. An independent noisy supervision panel B has 64 dates and 16 assets per date, with idiosyncratic noise standard deviation 0.5 and

common-date standard deviation 0.1. Its RNG stream is independent of train, search-set and holdout and is shared across holdout sizes. B supervises and tunes the selector; it never enters the predictor vector. The conditional expectation of this noisy label is population gain. The final auditor instead evaluates utility on an independent noiseless design with 2048 rows; its finite-sample utility need not equal that population quantity.

Target seeds 493600–493663 cross four families (linear, absolute interaction, maximum interaction and null), four worlds, holdout sizes $h \in \{2, 8, 16, 64\}$ and hill-climbing/evolutionary search. No target B is generated. For each held-out family and each h, source fitting and tuning use only the other three families. Family ID is never a feature. The folds test transfer to a specified excluded family. They do not provide a rule for choosing the appropriate selector when a deployment task's family is unknown.

Ordinary learned rules. Let s_i, v_i be search-set and holdout gains for shortlisted candidate i. Pooled selection scores $(16s_i + hv_i)/(16 + h)$. The convex blend tunes its search-set weight over $0, 0.1, \ldots, 1$ plus $16/(16 + h)$ on source-tune selected B utility. Affine ridge uses (s_i, v_i). Contextual ridge adds seven context variables: normalized search-set rank, first-proposal step and tree-node count, and the pool mean and sample standard deviation of each score. It also includes the fourteen products of each score with each context variable, giving 23 predictors. Source-fit means and scales standardize columns; the intercept is unpenalized and coefficient penalty is $n\lambda$, $\lambda \in \{0, 10^{-4}, 10^{-2}, 1\}$. Source-tune selected B utility chooses the penalty; the model is not refitted afterwards. Each rule competes against an explicit zero predictor with score and utility zero; ties retain the fixed shortlist order, with a nonzero candidate accepted only at a strictly positive maximum score. No source auditor utility enters fitting or tuning.

Inference and the failed gate. The six primary contrasts are contextual minus pooled, tuned blend and affine ridge under each target searcher, averaged equally over h, family and world within seed. The bootstrap jointly resamples the 64 seeds 20,000 times; Bonferroni marginal intervals cover 99.1667% each. They are approximate bootstrap intervals conditional on frozen source models, not a finite-sample guarantee or an assessment of source-training variability. The prespecified success criterion also requires every family-average contrast against affine ridge to be nonnegative. Table 14 retains every family: the absolute-interaction row fails this mean criterion even though its ordinary interval crosses zero. We retain the original criterion and report every family.

Table 14: All held-out families: contextual minus affine ridge, in 10^{-4} utility units. These secondary intervals are ordinary 95% seed-bootstrap intervals. Negative means in the absolute-interaction family make the original full gate fail.

Target search	Held-out family	Difference	Interval
Hillclimb	Linear	+0.980	[+0.302, +1.663]
Hillclimb	Absolute interaction	-1.065	[-2.598, +0.399]
Hillclimb	Maximum interaction	+0.534	[-0.758, +1.844]
Hillclimb	Null	+5.227	[+4.722, +5.735]
Evolutionary	Linear	+2.338	[+1.560, +3.136]
Evolutionary	Absolute interaction	-0.193	[-2.276, +1.882]
Evolutionary	Maximum interaction	+2.169	[+0.544, +3.884]
Evolutionary	Null	+5.471	[+4.987, +5.965]

Post hoc rejection-only controls. After the transfer result was inspected, two new source-trained rules retained the pooled nonzero winner and could only additionally reject it. Episodes rejected by pooled remain rejected. A scalar threshold is tuned on full source-tune B utility, using thresholds fixed by source-fit nonzero pooled-winner score quantiles (0 to 1 in steps of 0.05), clipped at zero, plus zero and all-reject. A six-feature ridge gate uses the pooled winner's two scores and four pool summaries, fits only source-fit episodes that pooled would accept, and tunes on all source-tune episodes including original zeros. It shares the original penalty grid and leave-family-out/source splits. Their models and choices were sealed again before this run's truth decoding, but previous target exposure still makes the comparisons post hoc.

Table 15: Post hoc contextual-minus-rejection-control contrasts, in 10^{-4} utility units. 98.75% marginal intervals adjust the four specified contrasts, conditional on frozen source models. Adjustment does not undo post hoc design or make this a fresh confirmation.

Target search	Rejection-only control	Difference	Interval
Hillclimb	Scalar threshold	+2.062	[+1.434, +2.657]
Hillclimb	Six-feature gate	+1.241	[+0.679, +1.784]
Evolutionary	Scalar threshold	+2.763	[+1.930, +3.619]
Evolutionary	Six-feature gate	+2.147	[+1.285, +3.045]

Separating ranking and acceptance. Write the pooled and learned rules as nonzero winners (i_P, i_C) and acceptance decisions (a_P, a_C). The four combinations retain the original candidate pools and score information (Table 16). Swapping a_C onto i_P reproduces 92.8% and 60.2% of the aggregate improvement under hill-climbing and evolutionary search. This ratio describes the component swap; it does not identify a causal contribution. The hybrid still requires the original learned model. Threshold/gate controls use less expressive acceptance functions than the full context model; their failure to match it does not rule out ordinary nonlinear or equally expressive rejection learning. Choosing a different program ID also need not change utility: distinct expressions sometimes implement equivalent predictions.

Table 16: Post hoc component swap: mean noiseless utility with pooled or learned nonzero ranking and acceptance. These combinations do not change discovery quality.

Rule components	Hillclimb	Evolutionary
Pooled ranking, pooled acceptance	0.01491873	0.01523801
Pooled ranking, learned acceptance	0.01503928	0.01536953
Learned ranking, pooled acceptance	0.01494110	0.01534523
Learned ranking, learned acceptance	0.01504865	0.01545641

Cost and interpretation. Target policies share $120 + 8$ candidate-evaluation queries and $512 + 256 + 16h$ label rows per world. Source supervision adds 1,048,576 logical labels and 8,388,608 candidate–label row evaluations, with panels shared across h; these counts are not independent sample sizes. Learned arms share the additional source evidence but have different fitting costs, and fixed pooling needs no source learning. No end-to-end cost or amortization advantage is established. The study uses no model API calls or real financial outcomes. The study finds that an ordinary finite-evidence selector transfers on average under the specified folds, while failing the prespecified family criterion. Ridge calibration, abstention and component swapping are not presented as new algorithms, economic mechanisms or recursive self-improvement.