A RISK-BOUNDED GATE FOR SELF-IMPROVING AGENTS: SELECTING MODIFICATIONS AS A MEAN–VARIANCE KNAPSACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting agent modifications by estimated mean gain alone ignores variability and interactions. We study a gate that selects from a fixed pool using a mean–variance knapsack with pairwise effects and a cost budget. Our contribution is a synthetic failure analysis of its measurement assumptions, not a validated agent-improvement system. We show that subtracting a shared base score from independently executed singleton scores induces common-reference covariance; this need not identify correlated modification effects within a joint execution. With five runs and twelve modifications, the sample covariance also has rank at most four. On fifteen fixed Gaussian instances, mean-only selection has true variance inflation between 2.98 and 20.57. A gate constrained at the estimated base stream-mean standard deviation retains median 26% of the true gain, and its selected portfolio fails the joint target of no variance inflation and at least 90% gain retention on every instance. An exploratory ceiling sweep illustrates sensitivity without establishing a transferable threshold. Separately, exact enumeration benchmarks heuristic accuracy and computational work on ten instances: diversity-seeded search recovers the optimum on a majority of restarts in all ten, but exceeds the target evaluation budget on all five twelve-item instances and on one sixteen-item instance. We release corrected code, regenerated synthetic results, historical protocol locks and a correction record. No valid live LLM-agent or ALFWorld result is reported. The findings distinguish feasibility under a fitted surrogate from evidence that a selected modification portfolio is reliable.

1 INTRODUCTION

An agent system in production is rarely a single design. It is a base agent plus an accumulating set of modifications: a workflow memory (Wang et al., 2024), a reasoning-trace bank (Ouyang et al., 2025), a reflection loop (Shinn et al., 2023), a step filter, a cache, extra restarts. Each is justified by a benchmark gain, and a self-improving system adds them on its own. Two facts make the accumulation risky. Gains are not additive. And gains are noisy: Ye et al. (2026) re-evaluate two memory-based self-improving methods and find that they *increase* run-to-run variance in 71% of cases, that best-to-worst gaps across identical runs reach 10 points, and that a shuffled task order turns a $+1.5\%$ improvement into a -4.5% degradation. Choosing modifications by their reported mean gain is a decision on a biased and noisy estimate, with no control over the risk it adds. We study a *gate*: a rule that decides which of a fixed pool of candidate modifications to keep under a cost budget and a fitted variance constraint. The central question is whether the measurements support the risk interpretation assigned to that constraint.

We treat the gate as a portfolio problem. Each modification k has a cost c_k and, on a stream of tasks, a random effect; a portfolio $x \in \{0,1\}^M$ has expected gain $\mu^\top x$ plus interaction terms and a proposed risk surrogate $x^\top S x$. Selecting x to maximise expected gain under a cost budget is a knapsack; adding a variance ceiling makes it a mean–variance knapsack (Markowitz, 1952; Goyal & Ravi, 2010). The mathematics is elementary. Its value is in (a) requiring μ, S, c to be *measured*, (b) exposing interactions that make naive stacking fail, and (c) giving a one-parameter knob $\sigma_{\max}$

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.

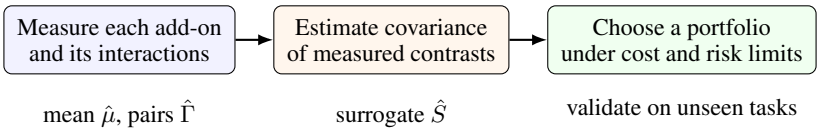

Figure 1: The proposed gate turns candidate agent modifications into a measured selection decision. Correlated fluctuations matter: the risk of a stack depends on how its add-ons fail together, not just on each add-on's variance. The covariance target must be checked before interpreting this as a risk bound.

that exposes the tradeoff between fitted gain and fitted risk. The harder part is measurement, and the paper is mostly about that.

Contributions and scope.

1. A mean–variance selection formulation with pairwise effects and phase costs, together with a derivation showing why separately measured singleton contrasts need not identify the covariance required by the structural risk model (Section 3).
2. A reproducible small-pool solver comparison against exact enumeration, reporting recovery and actual computational work separately (Section 4, Appendix C).
3. A synthetic stress test of the fitted gate: the joint variance and gain-retention target fails on all 15 instances. We distinguish target mismatch, finite-sample rank deficiency and exploratory ceiling sensitivity (Section 6).

All reported experiments are synthetic. The supplied live-agent and ALFWorld protocols are future evaluation plans, not evidence of practical benefit. We retain the historical hypotheses and disclose the code corrections that led to the current results.

2 Related Work

Self-improving agents and automated composition. Memory-based self-improvement Wang et al. (2024); Ouyang et al. (2025); Shinn et al. (2023), automated research agents Lu et al. (2024), evolutionary code search Novikov et al. (2025) and self-modifying agents Zhang et al. (2025a) generate candidate modifications faster than they can be validated; our work sits after generation. Yuan et al. (2025) select agentic components as a knapsack of success versus cost, without a variance ceiling, pairwise interactions or held-out variance inflation. ADAS (Hu et al., 2025) and AFlow (Zhang et al., 2025b) *search* over agent programs; AFlow uses run-to-run spread as a signal inside tree search. DSPy and MIPRO (Khattab et al., 2024; Opsahl-Ong et al., 2024) tune the components themselves. We do not invent or tune candidates: we take a frozen pool and treat variance as a fitted constraint audited on independent synthetic draws; a DSPy-compiled module would be one more item in the pool. Chen et al. (2024) show that even the simplest composition—more calls to the same model—is non-monotone in performance, which is why effects have to be measured rather than assumed additive.

Selecting from a library after measuring. The closest classical procedure is ensemble selection from a library of models (Caruana et al., 2004): greedy forward selection on a held-out metric, with bagging to control overfitting of the selection itself. Our estimation design plays the role of their library scores and our held-out stream of their test set; the differences are a hard budget, pairwise interactions, and a constraint on run-to-run variance rather than on the mean alone. "Portfolio" is used in the sense of Gomes & Selman (2001), who combined stochastic algorithms to cut the *variance* of run time; theirs is a mixture of independent runs, ours a conjunction of add-ons on one agent whose noise is correlated. Risk-averse Bayesian optimisation (Cakmak et al., 2020) optimises VaR or CVaR over a continuous design with a GP surrogate; our design is binary, the surrogate is a second-order model from a fixed design, and the risk measure is a hard ceiling.

Fragility, cost, and the classical pieces. Ye et al. (2026) document variance amplification and task-order dependence in self-improving agents and call for multi-run evaluation; the ceiling opera-

tionalises that call as a constraint. Multi-agent failure analyses include specification failures Cemri et al. (2025), and reported group advantages often vanish under matched compute Tran & Kiela (2026); Jwalapuram et al. (2026), which motivates including cost in the knapsack. The cost vector follows the construction / retrieval / generation phases of Omri et al. (2026). Mean–variance selection Markowitz (1952), chance-constrained knapsacks Goyal & Ravi (2010) and quality-diversity search Lehman & Stanley (2011); Mouret & Clune (2015) are classical; we claim no algorithmic novelty in the solvers.

3 Problem Formulation and Measurement

3.1 A structural model and a surrogate constraint

Let $x \in \{0,1\}^M$ select from a fixed pool of modifications. For a random task and execution, a second-order model is

$$Y(x) = b + \mu^\top x + x^\top \Gamma x + \xi + x^\top \eta, \tag{1}$$

where $b = \mathbb{E}[Y(0)]$, Γ is strictly upper-triangular, and ξ, η are zero-mean noises with $\mathrm{Var}(\xi) = \tau_0^2$, $\mathrm{Cov}(\eta) = \Sigma_{\text{eff}}$, and $\mathrm{Cov}(\eta, \xi) = s$. This is a structural approximation: omitted higher-order effects need not be zero-mean for each portfolio. The live packing score is bounded in $[0, 1]$; the Gaussian stress test in Section 6 is deliberately unbounded. Under this model,

$$\mathrm{Var}[Y(x)] = \tau_0^2 + 2x^\top s + x^\top \Sigma_{\text{eff}} x. \tag{2}$$

Negative base–effect covariance can offset added noise. A bound on the last term alone is therefore not a bound on total score variance.

For each item, $c_k = (c_k^{\text{con}}, c_k^{\text{ret}}, c_k^{\text{gen}})$ records construction, retrieval and generation costs (Omri et al., 2026); C collects these columns and λ converts them to common units. Given measured mean effects and interactions and a positive-semidefinite risk surrogate S, the proposed gate solves

$$\begin{aligned} \max_{x \in \{0,1\}^M} \quad & \hat{\mu}^\top x + x^\top \hat{\Gamma} x \\ \text{s.t.} \quad & \lambda^\top C x \le B, \qquad \sqrt{x^\top S x} \le \sigma_{\max}. \end{aligned} \tag{3}$$

Mean-only selection drops the second constraint. "Risk-bounded" here means feasible under this fitted surrogate; it is neither a deployment safety certificate nor an out-of-sample variance guarantee. Our experiments test the consequences of using the covariance estimator below for S.

3.2 What the estimators measure

The design contains the base, all singletons, all pairs and 30 random portfolios. Each is evaluated on n tasks in each of K runs. Write $\bar{Y}_r(x)$ for a run's stream mean and $\bar{Y}(x)$ for its mean over runs:

$$\begin{aligned} \hat{\mu}_k &= \bar{Y}(e_k) - \bar{Y}(0), \\ \hat{\gamma}_{kl} &= \bar{Y}(e_k + e_l) - \bar{Y}(e_k) - \bar{Y}(e_l) + \bar{Y}(0), \quad k < l, \\ g_{r,k} &= \bar{Y}_r(e_k) - \bar{Y}_r(0), \qquad \hat{S} = \widehat{\mathrm{Cov}}_r(g_r). \end{aligned} \tag{4}$$

Set $\hat{\Gamma}_{kl} = \hat{\gamma}_{kl}$ for $k < l$ and zero otherwise, so each interaction is counted once. The ceiling is scaled by $\hat{\sigma}_0 = \mathrm{sd}_r[\bar{Y}_r(0)]$, a *stream-mean* standard deviation, distinct from the single-score τ_0.

The distinction between $\hat{S}$ and Σ_{eff} is essential. Matching task order across separate calls does not couple their internal noise. The synthetic implementation draws fresh independent noise for every portfolio. Consequently its singleton contrasts estimate the covariance of *separately executed score differences*, rather than that of simultaneous add-on effects in one execution.

Proposition 3.1 (Common-reference covariance)**.** *Let $Z_0, Z_1, \ldots, Z_M$ be mutually independent stream means, with variances $v_0, v_1, \ldots, v_M$. For $g_k = Z_k - Z_0$,*

$$\mathrm{Cov}(g) = \mathrm{diag}(v_1, \ldots, v_M) + v_0 \mathbf{1}\mathbf{1}^\top. \tag{5}$$

In particular, its off-diagonal entries do not identify correlations between add-on noises within a joint execution.

Proof. Expanding $\mathrm{Cov}(Z_k - Z_0, Z_l - Z_0)$ gives v_0 for $k \neq l$ and $v_k + v_0$ for $k = l$, by independence. □

For the Gaussian generator below, singleton marginal variances are independent of ρ. Thus the population target of $\hat{S}$ does not identify ρ, even with infinitely many runs. Shrinkage cannot fix this target mismatch. Coupled counterfactual executions or an explicit fit to joint-portfolio variances would require a different measurement design; neither is evaluated here.

Independently of this mismatch, $\mathrm{rank}(\hat{S}) \leq \min(M, K-1)$. With $M = 12, K = 5$, its null space has dimension at least eight in $\mathbb{R}^{12}$; this does *not* imply that eight feasible binary portfolios lie in that space. We plot Ledoit–Wolf shrinkage $\hat{S}_\delta = (1-\delta)\hat{S} + \delta\bar{\lambda}I$ as a conditioning diagnostic (Ledoit & Wolf, 2004), not as proof of calibrated risk. For interactions we use paired run contrasts $d_{r,kl} = \bar{Y}_r(e_k + e_l) - \bar{Y}_r(e_k) - \bar{Y}_r(e_l) + \bar{Y}_r(0)$ and $\mathrm{SE}(\hat{\gamma}_{kl}) = \mathrm{sd}_r(d_{r,kl})/\sqrt{K}$. A 2 SE screen is descriptive, not a calibrated 5% test with five runs or a multiple-testing correction. LLM-estimated ROI values are not used as optimisation inputs.

3.3 Held-out endpoints

Definition 3.2 (Variance inflation)**.** For K held-out runs of n tasks and a fixed selected portfolio x_R,

$$\mathrm{VI}_{\mathrm{task}}(R) = \frac{\widehat{\mathrm{Var}}_{t,r}[Y_{t,r}(x_R)]}{\widehat{\mathrm{Var}}_{t,r}[Y_{t,r}(0)]}, \qquad \mathrm{VI}_{\mathrm{run}}(R) = \frac{\widehat{\mathrm{Var}}_{r}[\bar{Y}_r(x_R)]}{\widehat{\mathrm{Var}}_{r}[\bar{Y}_r(0)]}. \tag{6}$$

The first, the historical primary endpoint, pools task difficulty and execution variability; the second measures dispersion of stream means. They are not interchangeable. Retention is the held-out gain above base divided by the mean-only gain; it is undefined when that denominator is nonpositive. Degenerate base variance makes variance ratios uninformative. A solver satisfying (3) need not satisfy either VI threshold.

4 Solvers

Exact reference. For $M \leq 16$ we enumerate all 2^M portfolios, reject infeasible ones and maximise (3). This is exact for the supplied coefficients, not for unknown population effects. The mean-only and mean–variance rules use this oracle.

Variance-penalised greedy. At each step consider an affordable addition k and its score

$$s_\eta(k \mid x) = \hat{\mu}_k + \sum_{l : x_l = 1} (\hat{\Gamma}_{kl} + \hat{\Gamma}_{lk}) - \eta \left(\sqrt{(x + e_k)^\top S (x + e_k)} - \sqrt{x^\top S x} \right). \tag{7}$$

Add the highest positive-scoring item; if it violates the risk ceiling, undo it and stop. Single-start greedy uses $\eta = 0$. The corrected tuned baseline takes the best feasible result over $\eta \in \{0, .01, .03, .1, .3, 1, 3, 10, 30, 50\}$, charging all evaluations. This fixed grid replaces an ineffective historical bisection: the inner routine already enforced feasibility, so its feasibility test could not guide bisection. This correction is disclosed after the original lock; it is not a newly pre-registered comparison.

Diversity-seeded search and genetic baseline. Sample up to 400 portfolios to obtain up to eight feasible seeds with pairwise Jaccard index below 0.6. Improve each by add/drop/swap local search. From the best local optimum, try feasible improving additions of items found in other optima. The genetic baseline uses population 20, 15 generations, uniform crossover and per-bit mutation probability $1/M$, with a strict evaluation cap. Diversity search is *not* capped; its actual work must be compared with the target budget. No algorithmic novelty or equal-work superiority is claimed.

Solver check (N8-syn). Ten fixed generated instances, five each at $M = 12$ and 16, provide known coefficients and an exact optimum (Appendix C). Across 20 restarts, diversity search recovers that optimum on a majority of restarts for all ten instances; single-start greedy does so for four and the

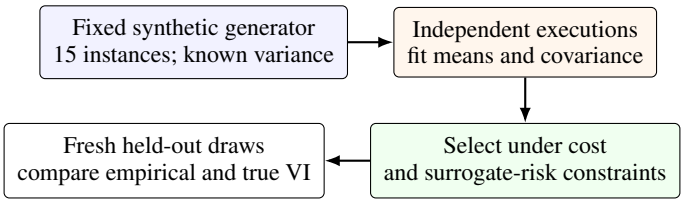

Figure 2: The experiment actually reported. All numbers concern synthetic data or exact solver checks. A separate live-agent design is supplied for future evaluation; it provides no empirical evidence in this submission.

GA for six. Diversity search uses mean 3,208 evaluations at $M = 12$ versus a 409 target; at $M = 16$ its overall mean is 5,850 versus 6,553, but one instance averages 6,606. These are descriptive accuracy/work tradeoffs, not a demonstrated hard-cap advantage. The original hypothesis fails at $M = 12$ and the per-instance budget condition also fails once at $M = 16$.

5 Experimental Design and Evidence Scope

Two distinct checks. N8-syn tests optimisation given known coefficients. N7-syn tests selection when coefficients are estimated from noisy data. Both use archived protocols dated 2026-09-11. The current results are corrected reruns after a code audit on 2026-09-26, with original generators, instance seeds, ceiling grid and hypothesis thresholds preserved. The audit corrected interaction double-counting, enforced the random baseline's cost budget, replaced the ineffective greedy tuning procedure, and made the GA cap strict. Original locks remain historical records; an appended correction notice and `CORRECTIONS.md` distinguish them from current results.

N7-syn design and rules. There are $109 = 1 + 12 + 66 + 30$ portfolios, each sampled for $K = 5$ runs of $n = 16$ scores: 545 estimation streams. Six rules are evaluated with five fresh held-out streams each: base; exact mean-only; exact mean–variance at $\sigma_{\max} = \hat{\sigma}_0$; tuned greedy; diversity search; and a random portfolio within the cost budget and within 25% of mean-only's cost (mean-only itself is the fallback if 200 draws fail). The later live protocol adds a seventh rule screening singleton effects at two standard errors; it is not part of the historical six-rule synthetic experiment. Costs are additive, $c_k \sim U(1, 10)$ and $B = 0.45 \sum_k c_k$.

Historical joint criterion. At $\rho = 0.8$, on at least four of five instances: true VI(mean-only) exceeds one, while the mean–variance portfolio has true VI at most one and true gain retention at least 90%. A secondary criterion compares median mean-only VI at $\rho = 0.8$ and 0. Failure of either rejects the joint hypothesis. This is a fixed finite-instance decision rule, not a population significance test. The sweep of ceilings $c\hat{\sigma}_0$, $c \in \{1, 1.5, 2, 3, 4\}$, is exploratory and uses held-out outcomes only for description, never for a confirmatory choice of c.

Unexecuted live evaluation. The supplied T1 protocol asks a frozen LLM to allocate 20 items to two agents under capacity constraints, scored by welfare divided by a dynamic programming optimum hidden from the LLM. It specifies 16 estimation tasks and 16 disjoint held-out tasks from two instance classes, and twelve prompt-level add-ons (Appendix B). These are analogues of memory, reflection, judging and search mechanisms, not reimplementations of the cited systems. A deterministic scorer removes judge noise but *not* task heterogeneity or parser effects. The protocol uses $\lambda = (0.1, 1, 1)$ for measured phase costs and the same 45% budget. Its historical two-model plan has no accepted result here. The later N7-v2 amendment specifies independent cache identities; it does not solve the covariance-identification problem in Proposition 3.1. No valid live-agent run is reported, and no live diagnostic CSV is shipped as experimental evidence.

Public-benchmark extension (T2). The released ALFWorld (Shridhar et al., 2021) pilot uses a ReAct-style agent (Yao et al., 2023), at most 30 actions per episode, six estimation and six held-out tasks, and three runs. It estimates singleton effects only ($\Gamma = 0$). This pilot has not been run and

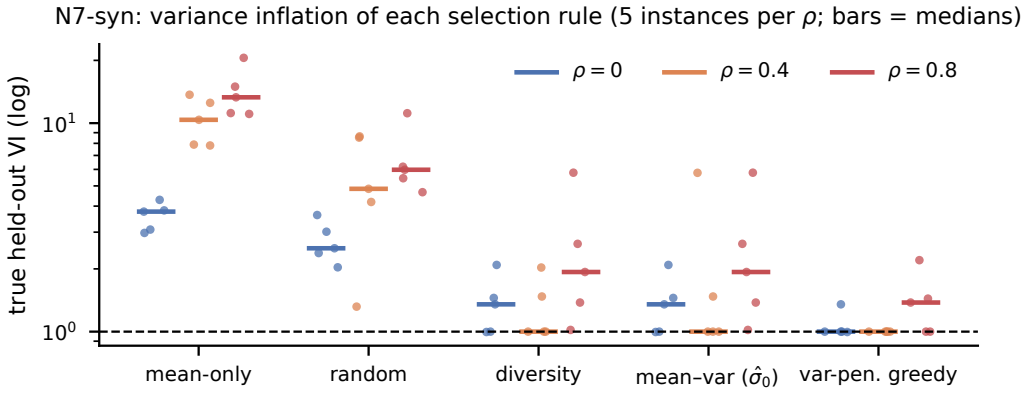

Figure 3: Corrected N7-syn results. Points are individual synthetic instances; bars are medians. True VI is analytic, not an LLM measurement. Variance-aware rules often reduce inflation but do not guarantee VI below one (dashed).

supplies no external validation. Its binary success score confounds mean and variance; Appendix G explains the additional descriptive endpoint.

Statistical interpretation. N7-syn reports analytic population VI and point estimates from 80 fresh held-out scores per rule. Its draws are independent and have no evolving memory or persistent task difficulty. It is therefore not a power study of task-order effects in agents. The future live analysis includes task bootstrap intervals (Efron, 1979); with dependent episodes these are conditional diagnostics, not validated confidence guarantees. Additional independent streams would be needed to study their coverage.

6 Synthetic Results

Generator. The frozen stress test uses

$$Y_{t,r}(x) = 0.7 + \mu^\top x + x^\top \Gamma x + (1 - a^\top x)\xi_{t,r} + x^\top \eta_{t,r}, \tag{8}$$

with independent $\xi \sim \mathcal{N}(0, \tau_0^2)$, $\tau_0 = 0.08$, and $\eta \sim \mathcal{N}(0, \Sigma_\eta)$. Here $\mu_k \sim U(-.01, .06)$, Γ_{kl} is $\mathcal{N}(0, .01^2)$ with probability .3 for $k < l$, $a_k \sim U(0, .15)$, and $\Sigma_\eta = DRD$ with $R = \rho \mathbf{1}\mathbf{1}^\top + (1 - \rho)I$ and $D_{kk} \sim U(.03, .08)$. There are five instances at each $\rho \in \{0, .4, .8\}$; different ρ values use different seeds for *all* parameters. Thus comparisons across ρ are descriptive, not paired interventions on correlation. Scores are not clipped to $[0, 1]$, preserving the analytic variance:

$$\mathrm{VI}_{\mathrm{true}}(x) = \frac{(1 - a^\top x)^2 \tau_0^2 + x^\top \Sigma_\eta x}{\tau_0^2}. \tag{9}$$

For this generator $\Sigma_{\mathrm{eff}} = aa^\top \tau_0^2 + \Sigma_\eta$; its stream-average counterpart is divided by $n = 16$. Each portfolio call draws new noise, so (5), not Σ_{eff}/n, is the population target of $\hat{S}$.

The joint target fails. Mean-only true VI ranges from 2.98 to 20.57; its medians are 3.76, 10.37 and 13.30 at increasing ρ. All 15 corresponding empirical point estimates exceed one. This demonstrates detection of large inflation on these particular iid instances, not that five runs suffice in general. The mean–variance rule retains median 26% of true gain and has median true VI 1.38 (Table 1); none of its five portfolios at any ρ satisfies the joint VI/retention target. It selects zero to four items (median two). The failure concerns the fitted gate and this generator; it does not prove that no other portfolio could meet the target, or that the same outcome would occur on an LLM agent. Random portfolios also inflate variance (median 4.67), but have lower median retention (0.72); this is not an equal-performance comparison.

Table 1: Corrected N7-syn: medians over 15 instances. Empirical VI uses 80 held-out scores per rule. Retention uses the true generator gain; base retention is omitted.

rule	true VI	empirical VI	retention
base	1.00	1.00	–
mean-only	10.37	10.03	1.00
mean–var ($\hat{\sigma}_0$)	1.38	1.47	0.26
penalty-grid greedy	1.00	1.03	0.00
diversity search	1.38	1.25	0.16
random (cost-matched)	4.67	3.92	0.72

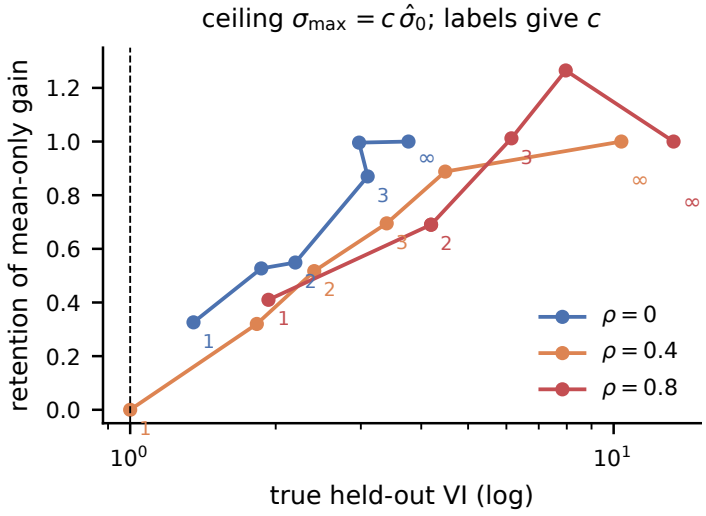

Figure 4: Exploratory ceiling sweep on corrected fits: coordinate-wise medians of true VI and true gain retention over five instances per ρ. Each point is a summary of different selected portfolios, not one attainable portfolio. No ceiling is selected using these outcomes.

Why a fitted bound can fail. Two issues coexist. First, independent singleton executions cannot identify within-portfolio noise correlation (Proposition 3.1). Second, the five-run estimate is rank deficient even for its own population target. Figure 8 separates the structural target, the covariance induced by the actual design, the sample estimate, and shrinkage. Regularising the sample matrix addresses conditioning, but neither the identification error nor the missing base–effect term in the total variance. We do not attribute the observed failure solely to rank deficiency or to an overly tight ceiling.

Exploratory sensitivity. At $c = 3$, median retention is 0.87, 0.69 and 1.01 across the three correlations, with median VI 3.10, 3.39 and 6.15. The last compares with mean-only median VI 13.30, but the ratio of these medians is not a median paired reduction. Retention can exceed one because mean-only optimises estimated, rather than true, gain. The non-monotone true frontier and uneven retention across settings show why $c = 3$ is not a general recommendation. Per-instance values and the full sweep are in Appendix D.

7 Discussion and Limitations

Practical implication. An exact optimiser can certify feasibility only for the matrix it receives. A useful modification gate must distinguish three checks: whether the coefficients represent the intended uncertainty; whether the solver satisfies its constraints; and whether selected portfolios meet held-out endpoints. Our solver experiment addresses the second. The synthetic experiment

exposes failures of the first and third. A well-conditioned matrix and an exact optimum alone do not establish reliable improvement.

What would make the gate usable? A follow-up design should estimate joint-portfolio variability, including the base–effect contribution, then validate a total-risk constraint on independent streams. Under the structural model, the existing base/singleton/pair design can identify a total-variance polynomial from portfolio variances (Appendix F); this alternative is not evaluated here. A common-random-number design may help when counterfactual executions can be meaningfully coupled; matching task order alone is insufficient. More runs can reduce covariance estimation error but cannot repair a wrong population target. Choosing a ceiling requires a separate validation split or a deployment risk specification, rather than the exploratory held-out sweep reported here.

Limits of the evidence. The score generator assumes the second-order mean structure, additive costs and Gaussian unbounded noise; it omits memory evolution, bounded-score saturation and higher-order stochastic interactions. Fifteen instances are a small stress test, not representative coverage of agent systems. Different seeds across correlation settings confound causal comparisons. The main endpoint pools task and run dispersion, whereas motivating agent studies often concern run-level success rates. N8-syn measures portfolio evaluations rather than wall-clock latency, and does not give diversity search and the GA equal work. Finally, there is no accepted LLM or ALFWorld experiment, no iterative self-improvement loop, and no evidence of improved deployment outcomes. The contribution is a measured failure analysis of a concrete gate, with a reproducible formulation and protocol.

8 Conclusion

Selecting agent modifications as a mean–variance knapsack makes cost and risk assumptions explicit. On the supplied synthetic stress test, mean-only selection inflates variance, but the fitted variance constraint fails its joint risk/retention target on all 15 instances. Independent singleton evaluations induce common-reference covariance and do not identify within-portfolio noise correlation; a five-run sample covariance adds rank deficiency. These findings delimit what the gate can certify. The released corrected code, data and historical protocols support reproduction of this negative result; live-agent effectiveness remains untested.

AI use statement

Generative AI coding and writing assistants were used strictly for drafting, literature search and editing text, as well as for LaTeX typesetting and figure rendering. AI assistance was not utilized for methodology design, mathematical review, or any code implementation, refactoring, and auditing.

Reproducibility statement

The supplement contains the synthetic generators, solvers, fixed seeds, corrected CSVs, table and figure generators, regression checks, and a machine-readable result manifest. Appendix H gives the commands and file mapping. Historical protocol locks are supplied with author names scrubbed; appended notices identify the 2026-09-26 post-lock corrections. No paid API call is required to reproduce any reported result. The proposed live harness and prompts are supplied as unevaluated research code; diagnostic live outputs are excluded.

References

Sait Cakmak, Raul Astudillo, Peter Frazier, and Enlu Zhou. Bayesian optimization of risk measures. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. arXiv:2007.05554.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004. doi: 10.1145/1015330.1015432.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. arXiv:2403.02419.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Carla P. Gomes and Bart Selman. Algorithm portfolios. *Artificial Intelligence*, 126(1–2):43–62, 2001. doi: 10.1016/S0004-3702(00)00081-3.

Vineet Goyal and R. Ravi. A PTAS for the chance-constrained knapsack problem with random item sizes. *Operations Research Letters*, 38(3):161–164, 2010.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2025.

Prathyusha Jwalapuram, Hehai Lin, Chuyuan Li, Fangkai Jiao, Sudong Wang, Yifei Ming, Zixuan Ke, Chengwei Qin, Giuseppe Carenini, and Shafiq Joty. The illusion of multi-agent advantage. *arXiv preprint arXiv:2606.13003*, 2026.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.03714.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

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

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.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Yasmine Omri, Ziyu Gan, Zachary Broveak, Robin Geens, Zexue He, Alex Pentland, Marian Verhelst, Tsachy Weissman, and Thierry Tambe. Agent memory: Characterization and system implications of stateful long-horizon workloads. *arXiv preprint arXiv:2606.06448*, 2026.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2406.11695.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.03768.

Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.

Qinyuan Ye, Yu Li, Yada Pruksachatkun, Jiaxin Zhang, and Chien-Sheng Wu. On the fragility of self-improving agents: Variance, task order, and underspecification. *arXiv preprint arXiv:2608.18066*, 2026.

Michelle Yuan, Khushbu Pahwa, Shuaichen Chang, Mustafa Kaba, Jiarong Jiang, Xiaofei Ma, Yi Zhang, and Monica Sunkara. Automated composition of agents: A knapsack approach for agentic component selection. *arXiv preprint arXiv:2510.16499*, 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*, 2025a.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2025b.

A NOTATION

symbol	meaning
x, M	binary modification portfolio and its size
μ, Γ	structural mean effects and strictly upper-triangular pair terms
$\tau_0^2, \Sigma_{\text{eff}}, s$	single-score base variance, effect covariance and base–effect covariance
$\hat{\sigma}_0$	sample standard deviation of base stream means
$S, \hat{S}$	risk surrogate and its estimate from singleton contrasts
K, n	runs and scores per run (5 and 16 in N7-syn)
ρ, a	generator noise correlation and base-noise attenuation coefficients
$\text{VI}_{\text{task}}, \text{VI}_{\text{run}}$	pooled-score and stream-mean variance ratios
retention	gain above base relative to mean-only gain; undefined if denominator is nonpositive

B THE FROZEN POOL

Table 2: The twelve prompt add-ons (`pool.md`, historical protocol), the mechanism each is an analogue of, and the intended cost profile (construction, retrieval, generation) in the units of Section 5. These are intended profiles, not validated live measurements. T2 uses the same twelve mechanisms with ALFWorld wording (`prompts/t2/mods/`).

id	add-on	what it changes in the packing agent (T1)	(con, ret, gen)
0	workflow memory	similar past (item, agent) assignments in context	(8, 4, 0)
1	trace memory	items that previously overflowed, listed as conflicts	(4, 4, 0)
2	reflection	same call: form a packing, revise once, return the revision	(0, 0, 20)
3	step judge	skip items below $0.8\times$ median remaining density	(0, 0, 2)
4	terminal judge	if slack and unpacked value are both high, retry in-call	(0, 0, 12)
5	reward cache	reuse cached density ranks from a similar fingerprint	(2, 6, 0)
6	top-k communication	each agent announces top-3 items; overlapping claims yield	(0, 0, 6)
7	one-bit contention	on a clash, the agent with more remaining capacity wins	(0, 0, 2)
8	random restart	three packing orders in one call; keep the best welfare	(0, 0, 24)
9	self-consistency	three independent calls; majority assignment	(0, 0, 24)
10	verifier	refuse overflow; harness also drops overflowing items	(0, 0, 2)
11	context compaction	catalogue truncated to the 8 densest items	(0, 0, 0)

C SOLVER EXPERIMENT: N8-SYN

The fixed generator draws $\mu_k \sim \mathcal{N}(.02, .03^2)$, pair terms $\Gamma_{kl} \sim \mathcal{N}(0, .01^2)$ with probability .3, costs $c_k \sim U(1, 10)$ and $B = .45 \sum_k c_k$. Noise standard deviations are $\sigma_k \sim U(.02, .06)$; a one-factor correlation matrix has loadings $\ell_k \sim U(.2, .8)$ and idiosyncratic variance $1 - \ell_k^2$. The ceiling is 1.5 times the mean σ_k. The instance seeds are 0–4 at each M, with 20 restarts per stochastic solver. Repeating the deterministic greedy result 20 times introduces no additional evidence.

Table 3: Corrected N8-syn. "Instances" counts majority recovery among 20 restarts. Evaluations and gain ratios are averaged over five instances. The cap is enforced only for the GA; diversity is assessed against it. One $M = 16$ diversity instance exceeds the cap despite the lower pooled mean.

M	solver	instances	restart hit	evals	cap	rel. gain
12	single-start greedy	3/5	0.60	31	409	0.865
12	penalty-grid greedy	4/5	0.80	235	409	0.939
12	GA (strict cap)	4/5	0.66	409	409	0.937
12	diversity search	5/5	0.99	3208	409	0.999
16	single-start greedy	1/5	0.20	42	6553	0.800
16	penalty-grid greedy	2/5	0.40	325	6553	0.896
16	GA (strict cap)	2/5	0.50	1116	6553	0.939
16	diversity search	5/5	0.99	5850	6553	0.998

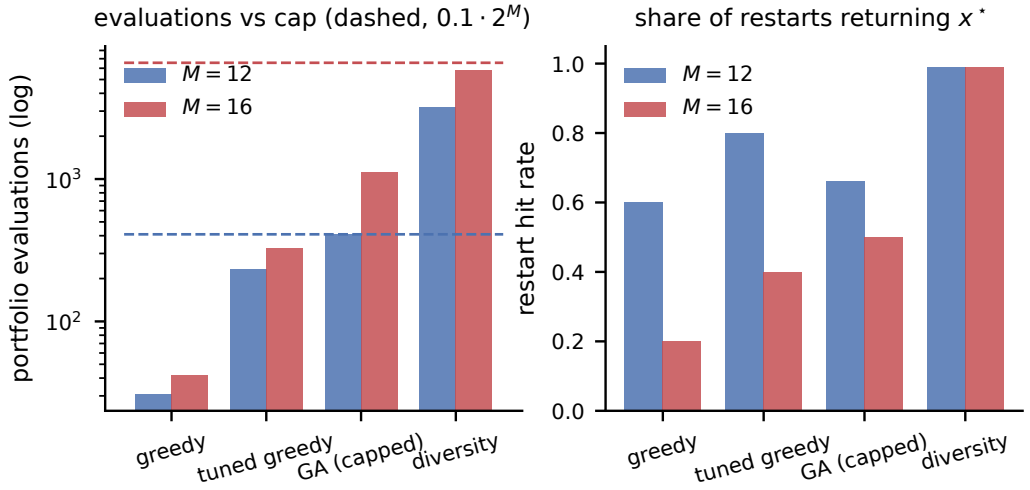

Figure 5: Solver recovery and work for N8-syn. Dashed lines mark target budgets; different solvers do not consume equal work.

Exact enumeration visits 4,096 or 65,536 portfolios. Diversity's mean 3,208 evaluations at $M = 12$ is below enumeration but far above the 409-evaluation target; the hypothesis required the latter. Its $M = 16$ mean of 5,850 is below the target 6,553, while seed 4 averages 6,605.6. These counts include candidate feasibility evaluations and repeated candidates; they are not timings or unique-portfolio counts.

D PER-INSTANCE RESULTS AND DIAGNOSTICS

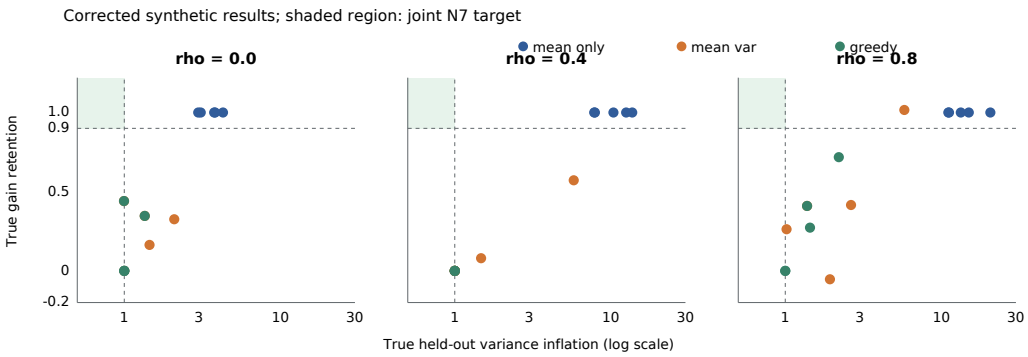

Figure 6: Corrected N7-syn portfolios. The shaded region is the joint target: true VI at most one and true retention at least 0.9. No selected mean–variance portfolio reaches it. This does not establish infeasibility over all possible portfolios.

Table 4: All corrected N7-syn instances. VI cells show true/empirical values; n is portfolio size. Retention is true gain of mean–variance (mv) over mean-only (mo). Rnd, div and greedy columns give true VI.

ρ	seed	VI mo	n	VI mv	n	ret. mv	VI rnd	VI div	VI greedy
0.0	0	4.29/6.81	6	1.00/1.47	0	0.00	2.04	1.00	1.00
0.0	1	3.09/2.17	6	1.35/1.50	2	0.35	2.38	1.35	1.35
0.0	2	2.98/2.40	4	1.00/0.96	1	0.44	2.51	1.00	1.00
0.0	3	3.76/5.26	7	2.09/2.68	2	0.33	3.62	2.09	1.00
0.0	4	3.81/4.41	7	1.45/1.35	2	0.16	3.02	1.45	1.00
0.4	0	10.37/10.03	6	1.00/1.10	0	0.00	8.53	1.00	1.00
0.4	1	12.52/13.47	7	5.78/5.97	4	0.57	4.84	2.03	1.00
0.4	2	13.68/9.26	6	1.00/0.62	0	0.00	4.19	1.00	1.00
0.4	3	7.81/5.49	4	1.00/0.70	0	0.00	1.32	1.00	1.00
0.4	4	7.89/10.22	6	1.47/2.08	1	0.08	8.64	1.47	1.00
0.8	0	14.96/18.16	6	2.64/3.17	2	0.42	5.98	2.64	1.44
0.8	1	13.30/19.25	6	5.79/4.20	4	1.02	4.67	5.79	2.20
0.8	2	11.07/10.10	6	1.02/1.16	1	0.26	5.44	1.02	1.00
0.8	3	11.18/11.41	5	1.93/1.82	2	-0.05	6.18	1.93	1.00
0.8	4	20.57/18.36	7	1.38/1.14	1	0.41	11.16	1.38	1.38

Table 5: Exploratory sweep: coordinate-wise medians (true VI, true retention), five instances per ρ. These are not paired reduction estimates.

c	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$
1.0	(1.35, 0.33)	(1.00, 0.00)	(1.93, 0.41)
1.5	(1.87, 0.53)	(1.83, 0.32)	(4.19, 0.69)
2.0	(2.20, 0.55)	(2.40, 0.52)	(4.19, 0.69)
3.0	(3.10, 0.87)	(3.39, 0.69)	(6.15, 1.01)
4.0	(2.98, 1.00)	(4.48, 0.89)	(7.96, 1.26)
∞	(3.76, 1.00)	(10.37, 1.00)	(13.30, 1.00)

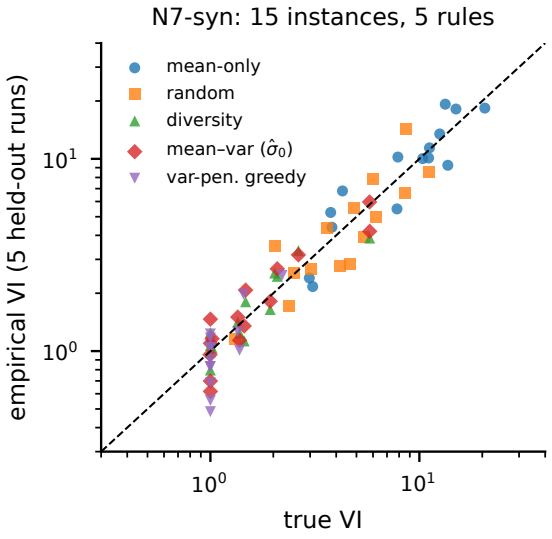

Figure 7: Empirical versus analytic true VI on corrected synthetic instances. Every mean-only point exceeds one on both axes; smaller departures near one can be missed. No detection-power guarantee is inferred.

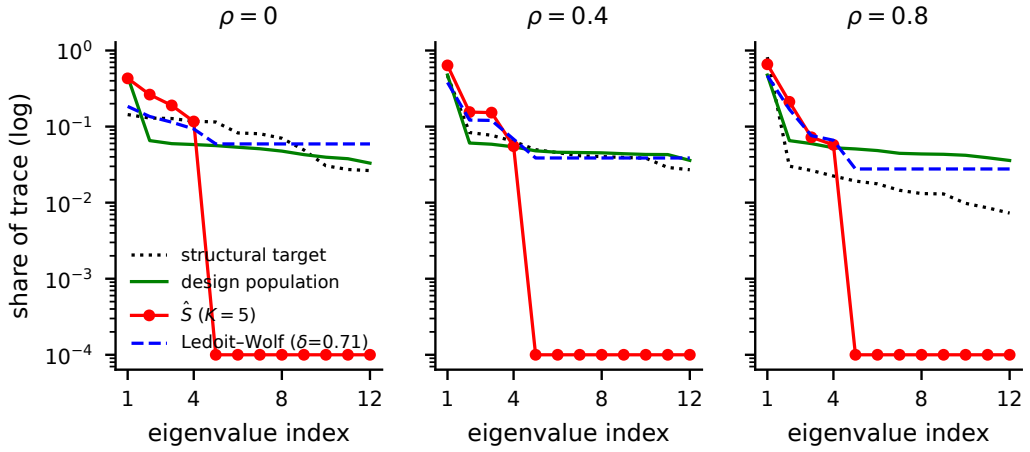

Figure 8: Seed 0 at each ρ: normalised eigenvalues of the structural stream-effect target, the population covariance induced by independent executions, the five-run sample estimate and its shrinkage. Each matrix is normalised by its own trace; this compares spectral shape, not absolute risk. Values below 10^{-4} are clipped for plotting, including numerical zero eigenvalues of the sample matrix.

E EVIDENCE BOUNDARIES

Table 6: What is and is not established.

layer	observed	scope
N8-syn	exact reference on ten generated instances; recovery/work tradeoffs	fixed-coefficient optimisation, not agent improvement
N7-syn	fifteen instances; joint target fails for every selected mv portfolio	behaviour of this generator, design and corrected pipeline
Live N7/T2	historical plans and harness; no accepted experiment	no empirical LLM or ALFWorld claim

The Gaussian generator is deliberately favourable to the second-order mean model, but it is not a model of task-order memory effects. Its latent noise correlation is known to the analyst, not identified by independent singleton measurements. This separation is the reason a correct solver and a deficient risk estimate can coexist.

F A POSSIBLE TOTAL-VARIANCE FIT (NOT EVALUATED)

Under the additive-noise model in (1), binary x makes $V(x) = \mathrm{Var}[Y(x)]$ a quadratic polynomial. Define

$$v_0 = V(0), \qquad d_k = V(e_k) - V(0), \tag{10}$$

$$d_{kl} = V(e_k + e_l) - V(e_k) - V(e_l) + V(0), \quad k < l. \tag{11}$$

Then $V(x) = v_0 + \sum_k d_k x_k + \sum_{k<l} d_{kl} x_k x_l$: expanding (2) gives $d_k = 2s_k + (\Sigma_{\mathrm{eff}})_{kk}$ and $d_{kl} = 2(\Sigma_{\mathrm{eff}})_{kl}$. Thus independent executions can identify total variance under this model if the variances of joint portfolios are used; the singleton-contrast covariance alone cannot. This identity suggests a repair using the already specified pair design. It is not an evaluated estimator: noisy fitted coefficients need not give nonnegative variances, and higher-order stochastic interactions invalidate the quadratic form. Estimation, regularisation and held-out calibration of such a gate remain future work. The regression checks verify only the algebra on the known generator, not statistical performance.

G BINARY-SCORE VARIANCE

In the unexecuted ALFWorld pilot, $Y \in \{0, 1\}$ and marginal variance is $p(1 - p)$. The proposed auxiliary statistic is

$$\mathrm{excess}(x) = \frac{\widehat{\mathrm{Var}}_r[\hat{p}_r(x)]}{\bar{p}(x)(1 - \bar{p}(x))/n}. \tag{12}$$

Its denominator is a plug-in iid Bernoulli reference. Task heterogeneity, dependent episodes and finite-sample estimation affect the ratio; it is not an isolation of agent-induced variance. It is undefined at a zero reference variance and is descriptive, not part of the N7 decision rule.

H REPRODUCTION AND CORRECTION RECORD

From `supplementary/code/`, install `requirements.txt`, then run:

```
python test_submission.py
python n7_synthetic.py
python n8_synthetic.py
python make_tables.py
python make_figures.py
python plot_tradeoff.py
```

The first two experiment scripts write the three synthetic CSVs in `results/`; `make_tables.py` writes all four numerical LaTeX tables and `results/submission_manifest.json` with hashes and summary checks. The figure scripts read those CSVs; the covariance spectrum reconstructs the fixed seed-0 designs. No local live result is automatically included in a synthetic figure. Runtime depends on the machine; all experiments are CPU-only and need no credentials.

The historical locks retain the original design and result narrative. Their appended correction notices and `CORRECTIONS.md` identify which results are superseded by the corrected CSVs. Hypothesis thresholds and instance generators were not retuned. Changes to the fitted interaction matrix and baselines are post-lock corrections, not evidence that the original implementation was correct. The supplement's `README.md` maps the unevaluated live harness and prompts separately.