
Calibrating the Search: Valid Inference for LLM-Driven Alpha Mining

Anonymous Author
Anonymous Institution

Abstract

Large language models (LLMs) are increasingly used to search for trading strategies: they propose formulas, backtest them, refine them using the results, and select the best one, so the same data both create and evaluate the reported winner. Corrections such as the Reality Check and the Deflated Sharpe Ratio adjust for this selection among the strategies tried; applied after an adaptive search, they treat that set as fixed, although backtest feedback produced it. We construct a simple adaptive search on data with no predictive signal and prove that three such corrections, from an adjustment for the number of trials to a Monte Carlo test that knows the true null and re-scores every logged formula, declare significance with probability tending to one as search depth and sample size grow. We propose generator-matched calibration (GMC), which reruns the entire search, feedback included, on datasets simulated from a fully specified no-signal model and compares the observed winner’s score with the winning scores of these reruns. Under a correctly specified null, GMC controls the reported winner’s Type-I error in finite samples. In a live LLM study, feedback raises that Monte Carlo test’s false-positive rate from 4.9% to 19.6%, while GMC’s estimated rate is 4.6% and its power is comparable to a holdout’s.

1 INTRODUCTION

In LLM-based strategy search, a researcher asks the model to propose a formula, backtests it, and feeds

the results back to guide the next proposal. Systems such as Alpha-GPT and AlphaAgent have made this a practical research workflow (Wang et al., 2025; Tang et al., 2025). The same historical data therefore both shape the strategies and evaluate them, so a winning backtest may reflect chance patterns that the search has learned to exploit rather than predictive signal.

Searching many strategies can produce attractive backtests even when none has predictive signal. White’s Reality Check (RC; White, 2000), Hansen’s Superior Predictive Ability (SPA) test (Hansen, 2005), and the Deflated Sharpe Ratio (DSR; Bailey and López de Prado, 2014) correct for this selection among the strategies tried. In an LLM feedback loop, however, backtest results also decide which strategies are tried at all. The set of tested strategies is then itself a function of the data, and correcting only for selection within it misses this second use of the data, which can undermine statistical validity (Dwork et al., 2015; Feng et al., 2022).

We show that recording every strategy tried and then applying such corrections does not fix the problem. We construct a simple adaptive search on data with no predictive signal and apply three corrections to its complete log, each using more information than the last. The strongest, the complete-log Monte Carlo test, knows the true null and re-scores every logged formula on fresh data drawn from it. We prove that all three declare significance with probability approaching one as the search depth and sample size grow. Because the Monte Carlo test is exact whenever the logged formulas do not depend on the data, its failure comes from one thing: holding fixed a log that the data helped generate.

To fix this, we propose generator-matched calibration (GMC), which calibrates the search procedure that produced the log (the *generator*), rather than the log itself. GMC asks how well the same search would do if there were no signal to find. It simulates independent datasets from a fully specified no-signal model, the global null, and reruns the same generator on each, so that feedback again generates new strategies and decides selection and stopping. A classical Monte Carlo

rank test (Dwass, 1957; Phipson and Smyth, 2010) then compares the observed winner’s score with the winners’ scores of these reruns (Figure 1).

A standard alternative is the holdout (sample splitting): the entire search, including any internal train-and-validation splits, runs on one part of the data, and its winner is tested once on the rest (Cox, 1975). A holdout stays valid under adaptive search but spends part of the sample on the final test; whereas GMC uses the whole sample both to search and to test.

In short, this paper asks:

How can we test whether the winner of an LLM-driven strategy search has real predictive signal, when the same data also guided the search?

Our contributions are:

1. **Failure of existing corrections.** We construct an adaptive search on data with no predictive signal on which an effective-count rule, the bootstraps of the Reality Check and SPA test, and the complete-log Monte Carlo test all declare significance with probability approaching one; the last fails even with a complete log and the true null. The full Deflated Sharpe Ratio instead over-rejects at practical depths and stops rejecting in the limit (Section 4).
2. **Generator-matched calibration.** We identify the complete search, including feedback, selection, and stopping, as the object to calibrate, and show that rerunning it on datasets from a correctly specified null controls the Type-I error of the reported winner in finite samples (Section 5).
3. **Evidence.** In simulations, the post-hoc corrections over-reject more as the search deepens while GMC does not, and GMC makes more true discoveries than a holdout at the prespecified setting. In a live LLM study, performance feedback raises the complete-log Monte Carlo test’s false-positive rate from 4.9% to 19.6%, while GMC’s estimated rate is 4.6% and its power is comparable to a holdout’s (Section 6).

2 RELATED WORK

Recent work evaluates the LLM discovery process as well as its outputs, through point-in-time reruns, separate feedback and test windows, and complete trial logs with search-aware corrections (Pan et al., 2026; Li et al., 2026; Gençay, 2026). Mehndiratta (2026)

notes that a candidate set that depends on past performance can violate the Reality Check’s assumption of a fixed candidate set, but its LLM experiments generate proposals without performance feedback, which matches the feedback-off condition of our live study (Section 6.2). Our theory and paired live experiment isolate the mechanism added by generating formulas from feedback. To our knowledge, ours is the first proof that complete-log Monte Carlo calibration can fail for feedback-driven strategy search, even when the null distribution is known exactly.

On the method side, Nikolopoulos (2026) calibrates diagnostics for financial machine-learning pipelines by repeating the full workflow in simulated environments; Appendix C.9 compares GMC with our reimplementations of that workflow. Closest to GMC, concurrent work by Li (2026) treats a complete LLM factor-selection pipeline as one selection rule, calibrates its reported Sharpe ratio by rerunning the pipeline on bootstrap null panels, and finds standard multiple-testing corrections miscalibrated under such search. Relative to both, we show why holding the generated strategies fixed breaks calibration and which parts of the search must be repeated for finite-sample Type-I error control. Reusable-holdout methods (Dwork et al., 2015) and methods for approving sequential model updates or selecting hyperparameters (Feng et al., 2022; Zecchin et al., 2025) protect validity under their own reuse protocols; GMC instead takes the search as given, and its rank test comes from classical Monte Carlo and randomization testing (Dwass, 1957; Phipson and Smyth, 2010; Hemerik and Goeman, 2018).

More broadly, the problem is selection followed by inference on the same data. Screening variables and refitting on the same data makes pure noise look significant (Freedman, 1983), and selecting features before cross-validation biases the error estimate unless the selection is repeated inside every resample (Ambroise and McLachlan, 2002); GMC applies that remedy to the whole search. Post-selection inference conditions on the selection event (Fithian et al., 2014), which is tractable when the event has a simple form, as in forward stepwise regression (Tibshirani et al., 2016), but not for an LLM search. Data fission and data thinning split each observation into independent parts and could replace a holdout that splits by date (Leiner et al., 2025; Neufeld et al., 2024), but they require a known noise family, such as Gaussian noise with known covariance; we compare GMC only with a holdout.

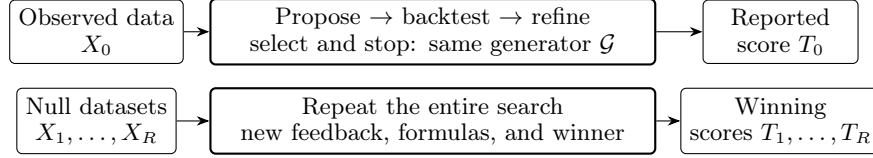


Figure 1: GMC reruns the whole search on each null dataset and compares the observed score T_0 with the winning scores $T_1, \dots, T_R$ of these reruns.

3 PRELIMINARIES

3.1 Problem Setup

Let X be all data available to a strategy search and ω its internal randomness. A generator $\mathcal{G}$ maps (X, ω) to an evaluation log $\mathcal{L}$ and a reported strategy W . Each new candidate may depend on the earlier formulas, their backtest scores, and any other feedback the search keeps, and the number of rounds may depend on this history too. The generator is fixed before inference, including its data access, feedback, allowed formulas, stopping rule, and handling of failed evaluations.

For a strategy w , write $f_w(t)$ for its return on date t . As a running example, we score a strategy by its studentized mean,

$$S(X, w) = \frac{\sqrt{n} \bar{f}_w}{s_n(w)}, \quad s_n^2(w) = \frac{1}{n} \sum_{t=1}^n (f_w(t) - \bar{f}_w)^2. \quad (1)$$

For fixed n , the annualized sample Sharpe ratio is a constant multiple of S . The search reports the score of the strategy it selects:

$$T(X, \omega) = S(X, W(X, \omega)). \quad (2)$$

The null distribution of T is determined by the whole procedure in (2), search and scoring together, rather than by the score of a formula treated as predetermined.

We test a *sharp global null*: a fully specified law P_0 for the data under which there is no predictive signal. Under this null, any rejection is a false positive, and we want $\mathbb{P}_0(\text{declare significance}) \leq \alpha$ for the one reported winner. This probability is over both the data and the search randomness, rather than conditional on the selected formula.

3.2 Existing Corrections

Existing corrections take one of two routes. The first, which we call a *fixed-output correction*, runs the search once and then corrects for selection among the formulas it produced, holding those formulas fixed. We analyze three such corrections, each using more information than the last: an *effective-count rule*, which raises

the significance threshold with an effective number of trials, much as a Šidák adjustment does with the total number; a *fixed-formula bootstrap*, which resamples the returns of the logged strategies and so keeps their dependence, much as the Reality Check does; and a *complete-log Monte Carlo test*, which knows the true null law and re-scores every logged formula on fresh datasets drawn from P_0 . All three see the complete log, and the last has no estimation or bootstrap error. The full Deflated Sharpe Ratio behaves differently, and Section 4 treats it separately.

The second route is the holdout, which stays valid because the search never sees its test data (Dwork et al., 2015).

We propose a third route, GMC (Section 5), which uses the full sample both to search and to test and, instead of holding the logged formulas fixed, reruns the complete search on each dataset drawn from P_0 (Figure 1).

4 COUNTEREXAMPLE

4.1 Construction

We construct a search on pure noise whose log grows only linearly in the search depth D but whose reported score grows like $\sqrt{D}$. Let $g_i(t)$, $i = 0, \dots, bD$, $t = 1, \dots, n$, be independent standard Gaussian variables, where $b \geq 2$ is fixed. For a unit vector $w \in \mathbb{R}^{bD}$, define

$$f_w(t) = \sqrt{\rho} g_0(t) + \sqrt{1-\rho} \sum_{i=1}^{bD} w_i g_i(t), \quad (3)$$

with fixed $\rho \in (0, 1)$. A strategy chosen without looking at the data has mean zero and variance one, and even a strategy chosen using the sample has zero expected return on new, independent data. The model thus has no predictive signal, even when the in-sample report looks significant.

Each coordinate $i \geq 1$ defines a basic strategy, the standard basis vector e_i , whose return mixes the common factor g_0 with its own noise g_i . The search runs for D rounds: it tries basic strategies in batches of b , keeps the best of each batch, and combines the noise

components of the kept strategies with equal weights, leaving the weight on the common factor unchanged.

Formally, partition the coordinates $1, \dots, bD$ into disjoint batches $B_1, \dots, B_D$ of size b , and write $A_i = \sqrt{n} \bar{g}_i$ for $i = 0, \dots, bD$, where $\bar{g}_i$ is the sample mean of g_i . In round d , the search scores the b basic strategies e_i , $i \in B_d$, and keeps the one with the largest sample mean return; because they share the common factor, its index is $i_d = \arg \max_{i \in B_d} A_i$. It then scores the *prefix* w_d , which weights the noise coordinates $i_1, \dots, i_d$ kept in rounds $1, \dots, d$ equally:

$$w_d = d^{-1/2} \sum_{j=1}^d e_{i_j}. \quad (4)$$

Because $i_1, \dots, i_d$ lie in disjoint batches, w_d is a unit vector. The generator reports $W = w_D$. Each i_j , and hence each w_d , depends on the data, so the evaluated combinations are not fixed in advance.

The complete log contains all bD tested basic strategies and all D evaluated prefixes:

$$\mathcal{L} = (e_1, \dots, e_{bD}, w_1, \dots, w_D), \quad K = (b+1)D.$$

Across possible datasets the final combination can take b^D forms, but any one search records only K evaluations. This gap between reachable combinations and recorded evaluations drives the separation in Theorem 1: the reported score grows like $\sqrt{D}$, whereas thresholds calibrated to a list of K evaluations grow only like $\sqrt{\log K}$.

Theorem 1 (Score inflation, informal). *Although there is no signal, the reported score grows like $\sqrt{D}$, so any test whose critical value grows more slowly rejects with probability tending to one.*

To see why, write the numerator as

$$\sqrt{n} \bar{f}_W = \sqrt{\rho} A_0 + \sqrt{1-\rho} D^{-1/2} \sum_{d=1}^D \max_{i \in B_d} A_i.$$

Each batch contributes a positive expected maximum, so the D selected maxima, weighted by $D^{-1/2}$, inflate the winner's score by an amount that grows like $\sqrt{D}$. Studentization does not remove this inflation: Gaussian sample means are independent of the centered data, so $ns_n^2(W)$ still has a χ_{n-1}^2 distribution after selection. Appendix A.1 states the theorem formally and proves it.

We use a Gaussian model because it makes the failure provable. The construction has two ingredients, both present when an LLM refines formulas using backtest feedback (Wang et al., 2025; Tang et al., 2025): the search keeps the candidates that score best in the sample, and later candidates build on them. Noise that

looks like signal then accumulates with depth, and a test that treats the selected formulas as fixed cannot see that selection.

4.2 Failure of Existing Corrections

Each of the three corrections of Section 3.2 fails on this search for the same reason: it holds fixed a log that the search generated from the data. Appendix A states each result formally and proves it.

Let $\hat{\Sigma}$ be the sample correlation matrix of *all* K logged return series. The effective-count rule uses its participation ratio as the number of trials:

$$\hat{N}_{\text{eff}} = \frac{(\text{tr } \hat{\Sigma})^2}{\|\hat{\Sigma}\|_F^2} = \frac{K^2}{\|\hat{\Sigma}\|_F^2}.$$

Because the basic strategies share the common factor in (3), their correlations alone keep $\|\hat{\Sigma}\|_F^2$ of order K^2 , whatever prefixes the search adds.

Proposition 2 (Effective-count rule, informal). *The effective number of trials stays bounded however deep the search, so the effective-count rule rejects with probability tending to one.*

Counting every logged evaluation as a separate test only raises the threshold to about $\sqrt{\log D}$, so a longer log does not help.

The fixed-formula bootstrap resamples the logged returns, which uses more information than a count but can still miss the same inflation. After the search, hold the generated strategies fixed and set $Y_{tj} = f_{v_j}(t)$ for their directions $v_1, \dots, v_K$. Let $y_{tj} = Y_{tj} - \bar{Y}_j$ and $s_j^2 = n^{-1} \sum_t y_{tj}^2$. Given the data, draw dates $I_1, \dots, I_n$ uniformly with replacement, the same for every strategy, and form

$$M^* = \max_{j \leq K} \frac{n^{-1/2} \sum_{r=1}^n y_{I_r j}}{s_j}. \quad (5)$$

Using the same dates for every strategy preserves the correlation between logged strategies on the same date. What it does not preserve is the dependence between the selected formulas and the data used to invent them.

Proposition 3 (Fixed-formula bootstrap, informal). *Bootstrap critical values that hold the logged formulas fixed grow only like $\sqrt{\log K}$, far below the $\sqrt{D}$ inflation, so these tests also reject with probability tending to one. This includes the block and stationary bootstraps used by the Reality Check and the SPA test.*

Resampling in blocks, dividing by a standard error that allows for autocorrelation, and recentering address serial dependence and poor strategies; none of them addresses the dependence between the log and the data.

The complete-log Monte Carlo test removes every approximation except the fixed formulas. It draws $X_1, \dots, X_R$ from P_0 , independently of each other and of (X, ω) , re-scores every logged formula on each, and ranks the observed best log score among the best log scores on the null datasets:

$$p_{\mathcal{L}} = \frac{1 + \sum_{r=1}^R \mathbf{1}\{T_{\mathcal{L}}(X_r) \geq T_{\mathcal{L}}(X)\}}{R + 1}, \quad (6)$$

where $T_{\mathcal{L}}(Y) = \max_{j \leq K} S(Y, v_j)$ is the largest score of the logged strategies on a dataset Y . It rejects when $p_{\mathcal{L}} \leq \alpha$.

Proposition 4 (Complete-log Monte Carlo test, informal). *The test is exact when the log does not depend on the data, yet on our construction it rejects with probability tending to one, even though it uses the true null.*

Exactness follows from the rank argument of Theorem 5 applied given the log, and covers any search that never sees the data it is tested on, such as an LLM without performance feedback. Feedback breaks this argument by making the log depend on the observed data: given the log, the observed data follow their conditional distribution given $\mathcal{L}$, whereas the fresh null datasets still follow P_0 . Knowing P_0 exactly does not recreate the selection that produced the log. On our construction the gap is large: the best log score on each null dataset grows more slowly than $\sqrt{D}$, whereas the observed best log score carries the inflation of order $\sqrt{D}$. Section 6.2 measures this contrast in a live LLM search.

Proposition 2 does not cover the full Deflated Sharpe Ratio, which also puts the standard deviation of the trial Sharpe ratios into its threshold (Bailey and López de Prado, 2014). This standard deviation grows as the search deepens, but too slowly to repair the test at practical depths: in simulations of our construction, counting every logged strategy as a trial, the DSR rejected in 5.2% of null replicates at $D = 4$, 13.6% at $D = 16$, and 77.8% at $D = 256$ (Appendix C.4). In the limit its threshold grows like $\sqrt{D \log D}$ and overtakes the inflation, so the DSR stops rejecting altogether, even against planted signals of fixed size (Proposition 6, Appendix A.6). Its false-positive rate is thus set by the search depth rather than by α .

5 METHOD

5.1 Generator-Matched Calibration

We therefore propose GMC, which uses the same p -value formula and null law as the complete-log Monte Carlo test (6), with one difference: each null dataset gets a fresh run of the complete search instead of

a re-scoring of the observed log. Fix a null model P_0 that we can simulate from, and the generator $\mathcal{G}$. Run the search on the observed data X_0 to obtain $T_0 = T(X_0, \omega_0)$; we call this the *target run*. Independently simulate $X_1, \dots, X_R$ from P_0 and run the same complete search on each with fresh internal randomness; we call these the *null runs*. Define

$$p_{\mathcal{G}} = \frac{1 + \sum_{r=1}^R \mathbf{1}\{T(X_r, \omega_r) \geq T_0\}}{R + 1}. \quad (7)$$

We reject when $p_{\mathcal{G}} \leq \alpha$. Null runs may choose different formulas, take different feedback paths, or stop at different times. These differences are exactly what a fixed log leaves out, and Proposition 4 shows that leaving them out makes the test fail.

Theorem 5 (Validity of GMC, informal). *If every null run repeats the same search procedure, GMC rejects a true null with probability at most α , whatever the procedure and however many null runs are used.*

The proof is the classical rank argument for exchangeable scores; Appendix B.1 gives it with the formal statement. It does not need a fixed number of candidates or independence between candidates, but it does need the same *procedure* in every run. For an LLM search this includes prompts, model configuration, feedback, parsing, retries, how invalid formulas are handled, and winner selection.

The number of null runs limits how small the p -value can be. At $\alpha = 0.05$, at least $R = 19$ null runs are needed to reject at all. With $R = 19$, the target must beat every null score; with $R = 99$, it must beat the fifth largest (Appendix B.3).

5.2 Practical Requirements

The search may be random; what matters is that the target and null runs are generated in exactly the same way, so that their scores are exchangeable under the null. With a correctly specified simulator, Theorem 5 applies directly, even when observations within a dataset are dependent.

Randomization gives another route when the data have a known symmetry, so GMC need not simulate a complete market model. Suppose a finite group H of transformations leaves the null distribution of the whole dataset unchanged, $\gamma X \stackrel{d}{=} X$ for every $\gamma \in H$; permuting returns across securities within each date is an example. Then the null runs can execute the complete search on randomly transformed copies of X , one independent, uniform draw from H per run, with i.i.d. search randomness, and the same rank bound holds, as in exact randomization testing (Hemerik and Goeman, 2018) (Appendix B).

A live LLM raises a further difficulty. Random responses are fine, but a shared model name or seed does not guarantee that all runs have the same distribution: provider updates, caching, shared state, or running target and null searches at different times can break exchangeability. Our live study therefore pins the model version, starts every search from a fresh context, interleaves all searches in a fixed run order, and checks GMC’s false-positive rate against separately generated sets of null searches (Section 6.2).

5.3 Comparison with a Holdout

A holdout and GMC may test different strategies: the holdout tests the winner of a search on part of the sample, whereas GMC tests the winner of a search on all of it, with no data withheld. Avoiding a split can improve power, because splitting discards information (Fithian et al., 2014), but it need not: GMC’s null scores come from the same adaptive search and are inflated too (Section 6.1).

6 EXPERIMENTS

We ask whether the failure appears in practice, whether GMC is calibrated, and whether it can still detect real signals, using two prespecified studies: a simulation matching Theorem 1 (Section 6.1) and a live study in which a production LLM searches real equity features (Section 6.2).

6.1 Simulation Study

The simulation runs the search of Section 4 with $n = 2520$ dates, $b = 8$ candidates per round, $\rho = 0.75$, and depths $D \in \{4, 8, 16\}$, so it tries up to 128 basic strategies, and reports the highest score in its log, which Theorem 1 also covers. Null A, our primary null, draws every $g_i(t)$ in (3) independently from $N(0, 1)$, and GMC simulates fresh null datasets from it. Null B adds serial dependence, for which the block and stationary bootstraps of Proposition 3 are designed: the dates form 120 independent 21-date blocks, and within a block each coordinate is Gaussian with correlation $0.5^{|r-s|}$ between dates r and s , independently across coordinates. Under Null B, GMC uses the randomization route of Section 5.2: each null run multiplies all data in each block by an independent random sign, which leaves the Null B distribution unchanged.

We compare GMC ($R = 19$ null runs) with a holdout that searches the first 84 of the 120 blocks and tests its selected strategy once on the last 36, and with two fixed-output corrections computed from the complete log (Appendix C.1): an *RC-style bootstrap* (the stationary bootstrap of Proposition 3, applied to the

largest average return without standardizing) and the effective-count rule (a participation ratio with a Šidák adjustment). All tests use level 5%, and each setting has 500 replicates.

To measure power, we plant a signal in the Null B design at depth 16: in each batch B_d of eight basic strategies, one strategy chosen at random receives the same positive mean return. The signal strength h is the expected z -statistic of the oracle strategy that combines the planted noise coordinates with equal weights; this oracle would reject with probability about 0.73 at $h = 2.25$ (Appendix C.2). Because a rejection alone does not show that the selected strategy carries the planted signal, we count a *true discovery* only when the test rejects and the selected strategy has positive population mean. So that extra discoveries are not simply extra false positives, the three comparators are recalibrated: each critical value becomes the 250th largest of its scores on 4,999 Null B replicates, an estimate of its 95th percentile. GMC keeps its own rank test.

Size. Under Null A, GMC rejects in 19, 15, and 18 of 500 replicates at depths 4, 8, and 16, and the holdout in 21, 19, and 19. GMC’s counts fall below 25 by chance: across 5,000 further replicates its rates are 5.2%, 4.9%, and 5.0% (Appendix C.1). As Section 4 predicts, the false-positive rate of the fixed-output corrections grows with depth: at depth 16, the effective-count rule rejects in 85.2% of null replicates and the RC-style bootstrap in 52.8% (Figure 2, left). Null B gives the same picture: GMC rejects in 23, 24, and 25 of 500 replicates, whereas at depth 16 the RC-style bootstrap still rejects in 58.6% (Table 1, Appendix C.1).

Power. On 500 paired replicates with the planted signal, GMC makes 71 true discoveries (14.2%), compared with 50 (10.0%) for the holdout (Figure 2, right; one-sided exact McNemar $p = 0.0167$). The recalibrated effective-count rule reaches 15.6%, so GMC improves on the holdout at this setting but not on the effective-count rule; the recalibrated RC-style bootstrap reaches 13.2% (Appendix C.2). Appendix C.3 reports other signal strengths and depths. The gain over the holdout depends on where the holdout splits the data: in an exploratory check with an even split, it shrinks to about one percentage point and is not significant (Appendix C.5).

6.2 Live LLM Study

The live study asks whether performance feedback in a real LLM search produces the failure of Proposition 4. We use DeepSeek-V4.1-Flash to run an eight-round search that requests 40 formulas per round in a fixed

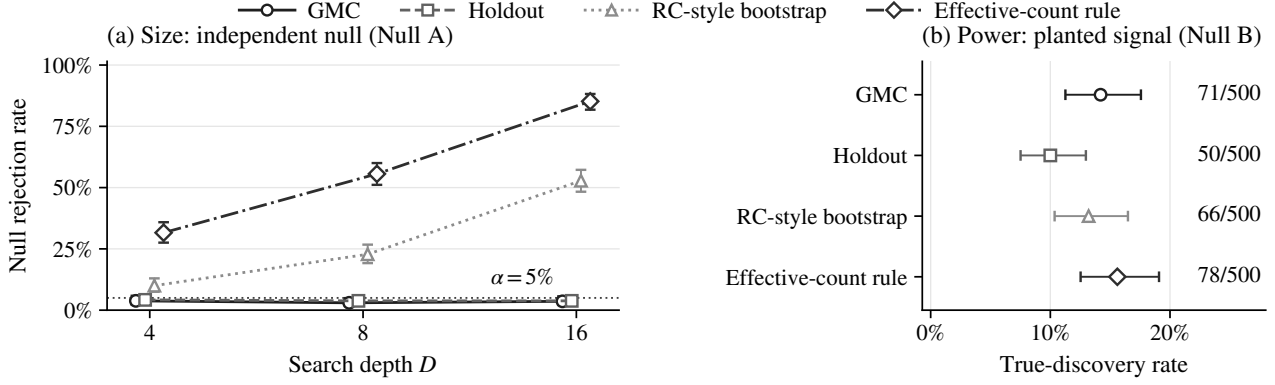


Figure 2: Simulation study (500 replicates per estimate; 95% Clopper–Pearson intervals). Left: false-positive rate at level 5% under Null A, which grows with search depth for the fixed-output corrections but not for GMC or the holdout (Null B: Table 1, Appendix C.1). Right: true-discovery rate under the planted signal ($D = 16$, $h = 2.25$), with the comparators recalibrated to level 5%.

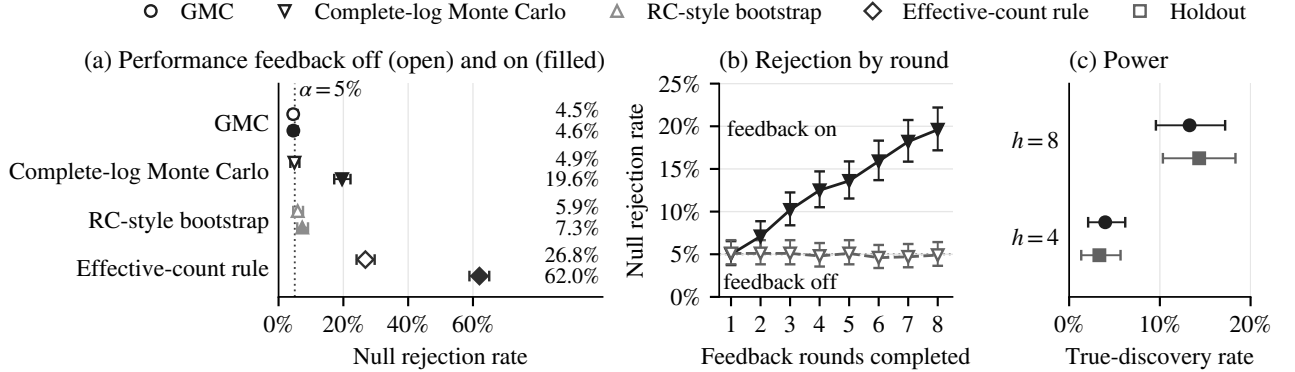


Figure 3: Live LLM study (95% intervals). (a) False-positive rates with feedback off (open markers) and on (filled), 1,000 searches each; GMC is evaluated, under each policy, against 19-run subsets of a separately generated set of 1,000 null searches (Appendix C.8); the holdout, valid by construction, is run only for the power study. (b) Complete-log Monte Carlo test after d rounds. (c) True-discovery rates at signal strength h , with GMC averaged over 19-run subsets of the 1,000 null searches with feedback.

expression language over 30 daily features of up to 499 large U.S. stocks. We hold this feature panel fixed, so X here denotes the simulated returns: independent standard Gaussian noise under the sharp null, which makes every rejection a false positive. Each formula defines a daily market-neutral portfolio, and each search reports the highest score in its complete log.

We compare two feedback policies on the same 1,000 null datasets. With feedback *on*, each prompt lists the 20 best-scoring formulas so far with their scores; with feedback *off*, it lists the 20 most recent formulas without scores. Everything else is identical. Without feedback, the log does not depend on the returns, so by Proposition 4 the complete-log Monte Carlo test is exact, and the off arm is a control whose false-positive

rate is known to be 5%.

Performance feedback quadruples the false-positive rate (Figure 3a). Without feedback, the complete-log Monte Carlo test rejects in 49 of the 1,000 searches (4.9%, 95% interval [3.6, 6.4]), as Proposition 4 predicts; with feedback, it rejects in 196 (19.6%, [17.2, 22.2]). The paired difference is 14.7 percentage points (conservative 95% interval [10.8, 18.5]), and both prespecified one-sided tests reject: of a rate above 5% with feedback, and of a higher rate with feedback than without (Appendix C.8).

The excess builds up round by round: when each search is cut off after d rounds, the rate with feedback rises from 5.0% after one round to 19.6% after eight, whereas without feedback it stays near 5% (Fig-

ure 3b). More null draws do not remove it: with $R = 999$, the rates are 23.4% and 4.8%. Figure 3a also shows the two other rules: the effective-count rule over-rejects even without feedback, and the RC-style bootstrap changes little. Appendix C.8 reports score summaries that show the mechanism.

We also estimate GMC’s own false-positive rate. For each policy, we generate a separate set of 1,000 null searches and test that policy’s 1,000 searches above against 19-run subsets of it. GMC’s estimated false-positive rate is 4.5% without feedback and 4.6% with feedback (approximate 95% interval 3.4–5.9% in both cases), consistent with the nominal 5% (Figure 3a; Appendix C.8, Table 11).

For power, we plant a signal defined by a fixed formula at strengths $h = 4$ and $h = 8$, the expected z -statistic of the oracle portfolio, in 300 datasets each. Each target search, run with feedback, is paired with a holdout search that selects on the first half of the dates and tests on the second. Here a *true discovery* is a rejection whose selected portfolio has cosine similarity at least 0.25 with the planted one. Estimated by averaging over 19-run subsets of the 1,000 null searches with feedback (Appendix C.8), GMC’s true-discovery rate is 4.0% at $h = 4$ and 13.3% at $h = 8$, against 3.3% and 14.3% for the holdout (Figure 3c). The approximate 97.5% intervals for the differences, $[-2.6, 3.9]$ and $[-7.3, 5.3]$ percentage points, allow sizable differences in either direction, so this study does not distinguish the two methods. Power is limited by the search more than by the test: only 8.0% and 15.0% of the target searches select a formula meeting the similarity criterion.

Across both studies, the fixed-output corrections over-reject as predicted, while GMC’s false-positive rate stays near 5%: at or below it under both simulated nulls, and 4.5% without and 4.6% with feedback in the live study. GMC finds more true signals than the holdout in the simulation and about as many in the live study.

7 DISCUSSION

Feedback changes more than the number of strategies tested: it changes which strategies a test is applied to, so fixed-output corrections can fail even with a complete log and the true null law. Calibrating the complete generator restores finite-sample control under a correctly specified null: validity comes from re-running the same search, randomness included. When all the data should take part in generating strategies, GMC provides a valid test with explicit conditions and a finite-sample guarantee; it complements a holdout rather than replacing it.

Several limitations remain. A simulator must represent the null being tested; if it is misspecified, a rejection may reflect the misspecification rather than predictive signal. Our live study pairs real features with Gaussian returns and plants a synthetic formula, so it measures the detection of such a formula rather than real predictability, and a simulated market need not represent real markets. An external LLM may not produce exchangeable runs; pinning the model version and checking against separately generated null searches reduce but do not remove this risk, and a locally hosted open-weight model would remove the dependence on an external service. A valid global rejection certifies neither the selected strategy’s future performance nor a particular economic mechanism, and the guarantee for one reported strategy does not extend to strong familywise or false-discovery-rate control under mixed nulls. Power relative to a holdout was higher in the simulation, indistinguishable in the live study, and lower at weak signals in an exploratory financial study (Appendix C.6). More null runs cost more but give finer p -values and less variable decisions. Nulls estimated from market data and error control when several strategies are reported are natural next steps.

AI Use Statement

Generative AI assistants were used in experiments planning and implementation and interpretation of results. They also assisted with checking proofs, experimental sources, and consistency between claims and evidence, and with preparing figures and tables. An LLM is also an explicitly studied component of the supporting strategy-search experiment. AI assistance is not treated as independent verification of a claim; the authors are responsible for the final mathematical arguments, experimental reporting, references, and submission.

References

- Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002. doi:10.1073/pnas.102102699.
- David H. Bailey and Marcos López de Prado. The Deflated Sharpe Ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi:10.3905/jpm.2014.40.5.094.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.
- D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975. doi:10.1093/biomet/62.2.441.
- Meyer Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957. doi:10.1214/aoms/1177707045.
- 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. doi:10.1126/science.aaa9375.
- Jean Feng, Gene Pennello, Nicholas Petrick, Berkman Sahiner, Romain Pirracchio, and Alexej Gossman. Sequential algorithmic modification with test data reuse. In *Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 674–684, 2022. <https://proceedings.mlr.press/v180/feng22a.html>.
- William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv:1410.2597, 2014. <https://arxiv.org/abs/1410.2597>.
- David A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983. doi:10.1080/00031305.1983.10482729.
- Eray Gençay. What survives honest evaluation? Leakage-safe, search-aware assessment of LLM-driven trading strategy discovery. arXiv:2608.27734, 2026. <https://arxiv.org/abs/2608.27734>.
- Stefano Giglio, Yuan Liao, and Dacheng Xiu. Thousands of alpha tests. *The Review of Financial Studies*, 34(7):3456–3496, 2021. doi:10.1093/rfs/hhaa111.
- Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi:10.1198/073500105000000063.
- Campbell R. Harvey and Yan Liu. False (and missed) discoveries in financial economics. *The Journal of Finance*, 75(5):2503–2553, 2020. doi:10.1111/jofi.12951.
- Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018. doi:10.1007/s11749-017-0571-1.
- Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi:10.1214/aos/1176347265.
- Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi:10.1214/aos/1015957395.
- James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025. doi:10.1080/01621459.2023.2270748.
- Jason Li. No better than random: An agent-level reality check for LLM factor selection. SSRN working paper, 2026. doi:10.2139/ssrn.7190318.
- 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:2608.28632v2, 2026. <https://arxiv.org/abs/2608.28632v2>.
- Parv Mehndiratta. More strategies, same zero: Multiple testing against LLM-scale alpha search. SSRN working paper, 2026. doi:10.2139/ssrn.7257858.
- Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024. <https://jmlr.org/papers/v25/23-0446.html>.

- Sotirios D. Nikolopoulos. Spurious predictability in financial machine learning. arXiv:2604.15531, 2026. <https://arxiv.org/abs/2604.15531>.
- Yingjian Pan, Xiaowei Ding, and Kay Giesecke. Agentic empirical asset pricing: Methodological foundations. arXiv:2609.00731, 2026. <https://arxiv.org/abs/2609.00731>.
- Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), Article 39, 2010. doi:10.2202/1544-6115.1585.
- Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi:10.1080/01621459.1994.10476870.
- Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013. doi:10.1214/ECP.v18-2865.
- Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsun Zheng, Keze Wang, Jinrui Chen, and Liang Lin. AlphaAgent: LLM-driven alpha mining with regularized exploration to counteract alpha decay. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 2813–2822, 2025. doi:10.1145/3711896.3736838.
- Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. doi:10.1080/01621459.2015.1108848.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi:10.1017/9781108231596.
- Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. In *Proceedings of EMNLP 2025: System Demonstrations*, pages 196–206, 2025. doi:10.18653/v1/2025.emnlp-demos.14.
- Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi:10.1111/1468-0262.00152.
- Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *International Conference on Machine Learning*, volume

267 of *Proceedings of Machine Learning Research*, pages 74018–74036, 2025. <https://proceedings.mlr.press/v267/zecchin25a.html>.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.** Sections 3–5 and Appendix C.
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes.** Theorems 1 and 5, Propositions 2–4, Section 5, and Appendices B.3 and C.9 state validity, resolution, and computational cost in units of complete search runs.
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **No.** Anonymized source code is not included with this submission; Appendix C describes each generator, null, and decision rule.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. **Yes.** The formal statements in Appendices A and B give the full assumptions; the main text states the results informally.
 - (b) Complete proofs of all theoretical results. **Yes.** All results are proved in Appendices A and B.
 - (c) Clear explanations of any assumptions. **Yes.** Section 5 distinguishes simulated nulls, joint invariance, and unverified live-provider assumptions.
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **No.** The paper gives the protocols and result counts, but the executable code, the financial panel, and the provider outputs are not included with this submission.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.** Appendix C describes the search, calibration, pilot selection, and held-out evaluation. The generator does not train the LLM.
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.** Section 6, captions, and Appendix C.
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.** Appendices C.8 and C.9 describe CPU execution and model inference through the provider’s API; the hardware model is not reported.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. **Yes.** The data and model providers are identified and linked in Appendices C.6 and C.8; methodological sources are cited.
 - (b) The license information of the assets, if applicable. **No.** Asset-by-asset license information is not included; the financial panel and provider outputs are not redistributed.
 - (c) New assets either in the supplemental material or as a URL, if applicable. **No.** No new assets are released with this submission.
 - (d) Information about consent from data providers/curators. **No.** No private participant data were collected; provider-specific permission and terms documentation are not included.
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.** The reported experiments concern numeric market features, synthetic returns, and strategy formulas, not personal participant records.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. **Not Applicable.** No human-subject experiment.
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

A PROOFS FOR THE COUNTEREXAMPLE

This appendix states the results of Section 4 formally and proves them; each subsection opens with the full statement, the informal version followed by the formal one. Section A.1 proves Theorem 1. By that theorem, a correction rejects with probability tending to one as soon as its critical value grows more slowly than $\sqrt{D}$, so the failure parts of Propositions 2–4 (Sections A.2–A.5) come down to bounding each correction’s critical value. Section A.6 treats the full Deflated Sharpe Ratio, whose critical value grows faster. Throughout, $b \geq 2$ and $\rho \in (0, 1)$ are fixed, and $n, D \rightarrow \infty$ along an arbitrary sequence unless a rate condition is stated.

A.1 Proof of Theorem 1

Theorem 1 (Score inflation). *Although there is no signal, the reported score grows like $\sqrt{D}$, so any test whose critical value grows more slowly rejects with probability tending to one.*

Formally, fix $b \geq 2$ and $\rho \in (0, 1)$ in (3)–(4), let $\mu_b = \mathbb{E} \max_{1 \leq i \leq b} Z_i > 0$ for independent $Z_i \sim N(0, 1)$, and let $n, D \rightarrow \infty$ jointly, at any relative rates. Then

$$\frac{S(X, W)}{\sqrt{D}} \xrightarrow{\mathbb{P}} c := \sqrt{1 - \rho} \mu_b > 0. \quad (8)$$

Consequently, a test that rejects when $S(X, W)$ exceeds a critical value $q_{n,D}$ growing more slowly than $\sqrt{D}$, in the sense that $(q_{n,D})_+ = o_{\mathbb{P}}(\sqrt{D})$, rejects with probability tending to one, even if $q_{n,D}$ depends on the data. The same holds if the reported score is the largest score in the complete log.

For each n , the variables $A_i = n^{-1/2} \sum_t g_i(t)$, $i = 0, \dots, bD$, are independent standard normals. Therefore the batch maxima $H_d = \max_{i \in B_d} A_i$ are i.i.d., with finite variance and mean μ_b , and their law does not depend on n . The mean is positive because $\max_{i \leq b} Z_i \geq \max(Z_1, Z_2)$ and $\mathbb{E} \max(Z_1, Z_2) = \mathbb{E}|Z_1 - Z_2|/2 > 0$. Since $W = D^{-1/2} \sum_d e_{i_d}$ and $A_{i_d} = H_d$,

$$\frac{\sqrt{n} \bar{f}_W}{\sqrt{D}} = \frac{\sqrt{\rho} A_0}{\sqrt{D}} + \sqrt{1 - \rho} \frac{1}{D} \sum_{d=1}^D H_d \xrightarrow{\mathbb{P}} \sqrt{1 - \rho} \mu_b$$

by the weak law of large numbers. The law of the left-hand side depends on D but not on n , so this convergence holds along any sequence with $D \rightarrow \infty$.

It remains to justify studentization after selection. Put $P_n = I_n - \mathbf{1}\mathbf{1}^\top/n$ and let $G_i = (g_i(1), \dots, g_i(n))^\top$. For a Gaussian sample, the sample mean is independent of the residual vector $P_n G_i$, and this holds jointly over all coordinates, so the collection $(P_n G_i)_{i=0}^{bD}$ is independent of $(A_i)_{i=0}^{bD}$. The selected W is a function of $(A_1, \dots, A_{bD})$ alone and has unit norm. Conditional on all A_i , therefore,

$$P_n f_W = \sqrt{\rho} P_n G_0 + \sqrt{1 - \rho} \sum_i W_i P_n G_i \sim N(0, P_n).$$

This conditional distribution depends neither on the means nor on the selected direction. Consequently $ns_n^2(W) = \|P_n f_W\|^2 \sim \chi_{n-1}^2$, independently of the numerator, and $s_n^2(W) \xrightarrow{\mathbb{P}} 1$ as $n \rightarrow \infty$; this law does not depend on D . Combining numerator and denominator proves (8) along any sequence with $n, D \rightarrow \infty$, without a relative growth condition.

For the rejection claim, fix $\epsilon \in (0, c)$. By (8), $\mathbb{P}\{S(X, W)/\sqrt{D} > \epsilon\} \rightarrow 1$, and by assumption $\mathbb{P}\{(q_{n,D})_+/\sqrt{D} < \epsilon\} \rightarrow 1$; on the intersection of these events, $S(X, W) > q_{n,D}$. If the reported statistic is instead the maximum studentized score over the complete log, it is at least $S(X, w_D)$, so the same argument applies; no monotonicity of the prefix scores is assumed. Finally, a convenient sufficient condition for $(q_{n,D})_+ = o_{\mathbb{P}}(\sqrt{D})$ is

$$q_{n,D} \leq C_{n,D} \sqrt{1 + \log m_D} + r_{n,D}, \quad (9)$$

with deterministic $m_D \geq 1$, $\log m_D = o(D)$, $C_{n,D} = O_{\mathbb{P}}(1)$, and $r_{n,D} = o_{\mathbb{P}}(\sqrt{D})$. It covers Gaussian-max thresholds over m_D tests, including the logged list with $m_D = K$. Indeed, (9) gives $(q_{n,D})_+ \leq |C_{n,D}| \sqrt{1 + \log m_D} + |r_{n,D}|$, and dividing by $\sqrt{D}$ yields $o_{\mathbb{P}}(1)$, because $C_{n,D} = O_{\mathbb{P}}(1)$, $\log m_D = o(D)$, and $r_{n,D} = o_{\mathbb{P}}(\sqrt{D})$.

Two remarks on the construction. Selection uses the raw numerator of each basic strategy, which has the same ordering as A_i , not a candidate-dependent studentized score. The log contains one duplicate because $w_1 = e_{i_1}$; removing it changes K to $K - 1$ and has no effect on the rates.

A.2 Proof of Proposition 2

Proposition 2 (Effective-count rule). *The effective number of trials stays bounded however deep the search, so the effective-count rule rejects with probability tending to one.*

Formally, under the assumptions of Theorem 1,

$$\hat{N}_{\text{eff}} \leq \left(\frac{b+1}{b\rho} \right)^2 + o_{\mathbb{P}}(1),$$

and for fixed $\alpha \in (0, 1)$, the rule that rejects when $S(X, W) > q_{n,D} := \Phi^{-1}((1 - \alpha)^{1/\max\{2, \hat{N}_{\text{eff}}\}})$ rejects with probability tending to one.

Write $M = bD$ and let r_{ij} be the sample correlation between the return columns f_{e_i} and f_{e_j} of two basic strategies. By the argument in the proof of Theorem 1, every logged column f_v satisfies $ns_n^2(v) \sim \chi_{n-1}^2$, so all empirical variances are positive almost surely for $n \geq 2$ and $\hat{\Sigma}$ is well defined. Conditional on the common-factor vector G_0 , the M columns of the basic strategies are i.i.d. Thus

$$U_{M,n} = \frac{1}{M(M-1)} \sum_{i \neq j} r_{ij}^2$$

is a conditional U-statistic of order two with kernel in $[0, 1]$. Terms with disjoint index pairs are conditionally independent and hence uncorrelated. At most $4M^3$ ordered pairs of index pairs share an index, and each such covariance is at most one in absolute value, against a normalization of order M^4 . Hence $\text{Var}(U_{M,n} | G_0) \leq C/M$ for a universal constant C , and the conditional Chebyshev inequality gives

$$\mathbb{P}\{|U_{M,n} - \mathbb{E}(r_{12}^2 | G_0)| > \epsilon\} \leq \frac{C}{M\epsilon^2} \quad \text{for every } n \text{ and } \epsilon > 0.$$

For a fixed pair of basic strategies, laws of large numbers for means, variances, and cross-products imply $r_{12} \rightarrow_{\mathbb{P}} \rho$ as $n \rightarrow \infty$. Because $r_{12}^2 \in [0, 1]$, also $r_{12}^2 \rightarrow \rho^2$ in L^1 , and conditional Jensen gives

$$\mathbb{E}|\mathbb{E}(r_{12}^2 | G_0) - \rho^2| \leq \mathbb{E}|r_{12}^2 - \rho^2| \rightarrow 0,$$

a bound that does not depend on D . The two displays together give $U_{M,n} \rightarrow_{\mathbb{P}} \rho^2$ along any sequence with $n, D \rightarrow \infty$. The full correlation matrix contains the principal submatrix of the basic strategies regardless of the selected prefixes, and therefore

$$\|\hat{\Sigma}\|_F^2 \geq M + M(M-1)U_{M,n} = M^2\{\rho^2 + o_{\mathbb{P}}(1)\}.$$

Since $\text{tr } \hat{\Sigma} = K = (b+1)D$ and $\rho > 0$,

$$\hat{N}_{\text{eff}} \leq \left(\frac{K}{M} \right)^2 \frac{1}{\rho^2 + o_{\mathbb{P}}(1)} = \left(\frac{b+1}{b\rho} \right)^2 + o_{\mathbb{P}}(1).$$

Dropping the duplicate changes K to $K - 1$ and leaves the same asymptotic bound. Finally, $N \mapsto \Phi^{-1}\{(1 - \alpha)^{1/\max(2, N)}\}$ is nondecreasing and finite, so with probability tending to one $q_{n,D} \leq \Phi^{-1}\{(1 - \alpha)^{1/\max(2, N_b+1)}\}$, where $N_b = \{(b+1)/(b\rho)\}^2$. Thus $(q_{n,D})_+ = O_{\mathbb{P}}(1) = o_{\mathbb{P}}(\sqrt{D})$, and Theorem 1 completes the argument.

A.3 Proof of Proposition 3(i)

Proposition 3 (Fixed-formula bootstrap). *Bootstrap critical values that hold the logged formulas fixed grow only like $\sqrt{\log K}$, far below the $\sqrt{D}$ inflation, so these tests also reject with probability tending to one. This includes the block and stationary bootstraps used by the Reality Check and the SPA test.*

Formally, assume the conditions of Theorem 1 and $D \log D = o(n)$. (i) Given the data, the $(1 - \alpha)$ quantile of (5) has positive part $O_{\mathbb{P}}(\sqrt{\log K})$, so a test comparing $S(X, W)$, or the largest score in the log, with this quantile rejects with probability tending to one. This remains true if each resample recomputes its denominators (Step 4 below) and if each bootstrap coordinate is shifted down by a data-dependent amount, as in the SPA test. (ii) The same holds for the two resampling schemes used with the Reality Check and the SPA test, the block bootstrap (Künsch, 1989), here in its circular form, and the stationary bootstrap (Politis and Romano, 1994), with or without studentization and recentering, provided the block length ℓ satisfies $\ell(D + \log n)(\log n)(\log D) = o(n)$; Appendix A.4 states the variants precisely.

Let $p = bD + 1$ and $X_t = (g_0(t), \dots, g_{bD}(t))^{\top}$. Each logged return is $a_j^{\top} X_t$ for the unit vector $a_j = (\sqrt{\rho}, \sqrt{1 - \rho} v_j^{\top})^{\top}$. The vectors a_j can depend on the full sample. Write

$$\hat{\Gamma} = \frac{1}{n} \sum_t (X_t - \bar{X})(X_t - \bar{X})^{\top}.$$

The proof has four steps. The tools are standard: the scalar Bernstein inequality, the multiplicative Chernoff lower-tail bound for sums of $[0, 1]$ -valued variables, and a net argument for operator norms (Boucheron et al., 2013; Vershynin, 2018).

Step 1: a high-probability event uniform over unit directions. For any fixed unit v , Gaussian quadratic-form concentration, or scalar Bernstein applied to $(v^{\top} X_t)^2 - 1$, gives

$$\mathbb{P} \left(\left| n^{-1} \sum_t (v^{\top} X_t)^2 - 1 \right| > z \right) \leq 2 \exp\{-\kappa n \min(z^2, z)\}$$

for a universal constant $\kappa > 0$. A $1/4$ -net of the unit sphere has at most 9^p elements, and $\|H\|_{\text{op}} \leq 2 \max_{v \text{ in net}} |v^{\top} H v|$ for every symmetric matrix H . A union bound over the net therefore shows that the uncentered second-moment matrix differs from I_p in operator norm by $O_{\mathbb{P}}(\sqrt{p/n} + p/n)$. Also $\|\bar{X}\|^2 = O_{\mathbb{P}}(p/n)$, and subtracting $\bar{X}\bar{X}^{\top}$ shows that $\hat{\Gamma}$ has the same rate. Gaussian norm concentration and a union bound over the n rows give $\max_t \|X_t - \bar{X}\|^2 = O_{\mathbb{P}}(p + \log n)$. Since $D \log D = o(n)$ implies $p/n \rightarrow 0$, for every $\varepsilon > 0$ there is a constant C_{ε} such that, for all large n , the event

$$E_n = \left\{ \|\hat{\Gamma} - I_p\|_{\text{op}} \leq \frac{1}{2}, \quad \max_t \|X_t - \bar{X}\|^2 \leq C_{\varepsilon}(p + \log n) \right\}$$

has probability at least $1 - \varepsilon$. The remaining steps work on E_n .

Step 2: bounding the largest standardized observation. On E_n , $\lambda_{\min}(\hat{\Gamma}) \geq 1/2$, so every logged column has $s_j^2 = a_j^{\top} \hat{\Gamma} a_j \geq 1/2$, whatever the data-dependent directions a_j . Set $x_{tj} = y_{tj}/s_j$ and $\eta_n = \max_{t,j} x_{tj}^2/n$. Because $y_{tj} = a_j^{\top} (X_t - \bar{X})$ and $\|a_j\| = 1$,

$$\eta_n \leq \frac{\max_t \|X_t - \bar{X}\|^2}{n \lambda_{\min}(\hat{\Gamma})} \leq \frac{2C_{\varepsilon}(p + \log n)}{n}.$$

This bound is uniform in the selected directions; no conditioning on a fixed candidate family is used. Outside E_n , if some $s_j = 0$, the critical value is defined as $+\infty$ and the test does not reject; since $\mathbb{P}(E_n) \geq 1 - \varepsilon$, this convention does not affect the limit.

Step 3: conditional bounds. Conditional on the data, put $A_j^* = n^{-1/2} \sum_r x_{I_r j}$ and $Q_j^* = n^{-1} \sum_r x_{I_r j}^2$. The summands of A_j^* are independent, centered because $\sum_t y_{tj} = 0$, have total variance one because $n^{-1} \sum_t x_{tj}^2 = 1$, and are bounded in absolute value by $\sqrt{\eta_n}$. Let $u = \log(6K/\alpha)$. Bernstein's inequality, in the form $\mathbb{P}(|S| > \sqrt{2vu} + \frac{2}{3}Lu) \leq 2e^{-u}$ for a sum S of independent centered summands with total variance v and absolute bound L , gives

$$\mathbb{P}^*(|A_j^*| > r_n) \leq 2e^{-u}, \quad r_n := \sqrt{2u} + \frac{2}{3}u\sqrt{\eta_n}.$$

The summands $x_{I_r j}^2/n$ of Q_j^* are independent, lie in $[0, \eta_n]$, and have total mean one. Applying the multiplicative Chernoff lower-tail bound to their rescaled versions, which lie in $[0, 1]$, gives

$$\mathbb{P}^*(Q_j^* < 1 - \delta_n) \leq e^{-u}, \quad \delta_n := \sqrt{2\eta_n u},$$

whenever $\delta_n \leq 1$. A union bound over the K columns shows that the event

$$E_n^* = \left\{ \max_j |A_j^*| \leq r_n, \quad \min_j Q_j^* \geq 1 - \delta_n \right\}$$

has conditional probability at least $1 - 3Ke^{-u} = 1 - \alpha/2$.

Step 4: rates and the two quantiles. On E_n , $\eta_n u \leq 2C_\varepsilon(p + \log n) \log(6K/\alpha)/n$. Here $p = bD + 1$ and $K = (b+1)D$, so $(p \log K)/n \rightarrow 0$ by $D \log D = o(n)$, and $(\log n)(\log K)/n \rightarrow 0$ because $D < n$ for all large n . Hence $\eta_n u \rightarrow 0$ uniformly on E_n , and $u/n \rightarrow 0$. In particular, for all large n , $\delta_n \leq 1$ and $r_n = \sqrt{u} \{\sqrt{2} + \frac{2}{3}(\eta_n u)^{1/2}\} \leq 2\sqrt{u}$.

For the fixed-denominator bootstrap, the statistic (5) equals $\max_j A_j^*$, which is at most $2\sqrt{u}$ on E_n^* . Since $\mathbb{P}^*(E_n^*) \geq 1 - \alpha/2 > 1 - \alpha$, the conditional $(1 - \alpha)$ quantile of (5) is at most $2\sqrt{u}$.

For the restudentized bootstrap, every denominator is recomputed from the resample: the squared denominator is $s_{j,*}^2 + n^{-2}s_j^2$, where $s_{j,*}^2$ is the resample variance defined with divisor n . Dividing it by s_j^2 gives exactly

$$Q_j^* - \frac{(A_j^*)^2}{n} + n^{-2}.$$

On E_n^* , $Q_j^* \geq 1 - \delta_n$ and $(A_j^*)^2/n \leq r_n^2/n \leq 4u/n$, so for all large n the squared denominator is at least $s_j^2/2$ for every j . The restudentized maximum is then at most $\sqrt{2} \max_j |A_j^*| \leq 2\sqrt{2u}$ on E_n^* , and its conditional $(1 - \alpha)$ quantile is at most $2\sqrt{2u}$. The constant differs from the fixed-denominator case, but the order $\sqrt{u}$ is the same. The ridge $n^{-2}s_j^2$ keeps the denominator positive even when a bootstrap sample repeats a single row.

In both cases, with probability at least $1 - \varepsilon$ for all large n , the positive part of the critical value is at most $2\sqrt{2 \log(6K/\alpha)}$. Since ε is arbitrary, it is $O_{\mathbb{P}}(\sqrt{\log K})$, and $\log K = o(D)$ gives $(q_{n,D})_+ = o_{\mathbb{P}}(\sqrt{D})$. Theorem 1 then yields rejection with probability tending to one, both for $S(X, W)$ and for the maximum logged score.

Recentring. A recentring such as that of the SPA test adds a nonpositive, data-dependent shift to each bootstrap coordinate. Lowering coordinates cannot raise their maximum, so the conditional $(1 - \alpha)$ quantile of the recentered maximum is at most that of the fully centered one, and the bounds above apply.

A.4 Proof of Proposition 3(ii)

Part (ii) of Proposition 3 asserts the following. Under its assumptions, the same conclusion holds for the circular block bootstrap with block length ℓ dividing n and for the stationary bootstrap with restart probability $1/\ell$, provided $\ell(D + \log n)(\log n)(\log D) = o(n)$, whether the maximum is taken over raw centered means or over means studentized by the observed Bartlett long-run variance with bandwidth ℓ , with or without recentring, each compared with its observed counterpart. We now prove this.

Keep the notation of Appendix A.3: the $X_t = (g_0(t), \dots, g_{bD}(t))^\top$ are independent $N(0, I_p)$ vectors with $p = bD + 1$, every logged return is $a_j^\top X_t$ for a possibly data-dependent unit vector a_j , and $\tilde{X}_t = X_t - \bar{X}$. Indices are taken modulo n . For $1 \leq r \leq n$ define the centered circular block sums and block covariance matrices

$$S_{s,r} = \sum_{a=0}^{r-1} \tilde{X}_{s+a}, \quad \hat{\Omega}_r = \frac{1}{nr} \sum_{s=1}^n S_{s,r} S_{s,r}^\top.$$

The long-run variance of column j is $\hat{\omega}_j^2 = a_j^\top \hat{\Omega}_\ell a_j$, the Bartlett estimate with bandwidth ℓ computed from circular autocovariances. Put $L = \ell$ for the block bootstrap and $L = \min\{n, \lceil 1 + \ell \log(2n/\alpha) \rceil\}$ for the stationary bootstrap. In both cases the rate condition gives $L\{p + \log(nL)\} \log K = o(n)$, because $L = O(\ell \log n)$, $p + \log(nL) = O(D + \log n)$, and $\log K = O(\log D)$.

Step 1: uniform bounds for block covariance matrices and block sums. We show that

$$\max_{r \leq L} \|\hat{\Omega}_r - I_p\|_{\text{op}} = o_{\mathbb{P}}(1), \quad \max_{s \leq n, r \leq L} \frac{\|S_{s,r}\|^2}{r} = O_{\mathbb{P}}\{p + \log(nL)\}.$$

Fix $r \leq L$ and a unit vector u , and put $\eta_t = u^\top X_t$; these are independent standard normal variables. Without the centering, the matrix $\hat{\Omega}_r^0 = (nr)^{-1} \sum_s (\sum_{a < r} X_{s+a})(\sum_{a < r} X_{s+a})^\top$ satisfies $u^\top \hat{\Omega}_r^0 u = \eta^\top M_r \eta$, where $M_r =$

$(nr)^{-1}B_r^\top B_r$ and B_r is the $n \times n$ circulant matrix that maps a series to its n circular block sums of length r . Each row and each column of B_r contains r ones, so $\|B_r\|_{\text{op}} = r$. Hence $\|M_r\|_{\text{op}} \leq r/n$, $\text{tr } M_r = 1$, and $\|M_r\|_F^2 \leq \|M_r\|_{\text{op}} \text{tr } M_r \leq r/n$. The Hanson–Wright inequality (Rudelson and Vershynin, 2013) gives, for a universal constant $\kappa > 0$,

$$\mathbb{P}\{|u^\top \widehat{\Omega}_r^0 u - 1| > z\} \leq 2 \exp\{-\kappa n \min(z^2, z)/r\}.$$

A union bound over a $1/4$ -net of the unit sphere, with at most 9^p points, and over $r \leq L$, together with $\|H\|_{\text{op}} \leq 2 \max_{v \text{ in the net}} |v^\top H v|$ for symmetric H , gives $\max_{r \leq L} \|\widehat{\Omega}_r^0 - I_p\|_{\text{op}} = O_{\mathbb{P}}[\{L(p + \log L)/n\}^{1/2}] = o_{\mathbb{P}}(1)$. Circular counting gives the exact identity $\widehat{\Omega}_r = \widehat{\Omega}_r^0 - r\bar{X}\bar{X}^\top$, and $r\|\bar{X}\|^2 \leq L\|\bar{X}\|^2 = O_{\mathbb{P}}(Lp/n) = o_{\mathbb{P}}(1)$, which proves the first bound. For the second, $S_{s,r}$ is a combination of the X_t with coefficients $\mathbf{1}\{t \text{ lies in the block}\} - r/n$, whose squares sum to $r(1 - r/n)$. Therefore $\|S_{s,r}\|^2/\{r(1 - r/n)\}$ has a χ_p^2 distribution, and the bound $\mathbb{P}\{\chi_p^2 \geq p + 2\sqrt{px} + 2x\} \leq e^{-x}$ (Laurent and Massart, 2000), applied with $x = \log(nL) + y$ and a union bound over the nL pairs (s, r) , proves the second.

Step 2: the circular block bootstrap. Let $n = m\ell$ and put $Z_{sj} = \ell^{-1/2}a_j^\top S_{s,\ell}$, so that $n^{-1}\sum_s Z_{sj} = 0$ and $n^{-1}\sum_s Z_{sj}^2 = \widehat{\omega}_j^2$. Given the data, the bootstrap draws block starts $J_1, \dots, J_m$ independently and uniformly from $\{1, \dots, n\}$ and concatenates the blocks. The bootstrap mean of column j , minus $\bar{Y}_j$ and multiplied by $\sqrt{n}$, equals $m^{-1/2}\sum_q Z_{J_q j} = \widehat{\omega}_j A_j$, where $A_j = m^{-1/2}\sum_q x_{J_q j}$ and $x_{sj} = Z_{sj}/\widehat{\omega}_j$. Conditionally, the summands $x_{J_q j}/\sqrt{m}$ are independent, centered, have total variance one, and are bounded in absolute value by $\sqrt{h}$, where by Step 1

$$h = \max_{s,j} \frac{x_{sj}^2}{m} \leq \frac{\max_s \|S_{s,\ell}\|^2/\ell}{m \lambda_{\min}(\widehat{\Omega}_\ell)} = O_{\mathbb{P}}\left\{\frac{\ell(p + \log n)}{n}\right\}.$$

With $u = \log(4K/\alpha)$, Bernstein’s inequality in the form used in Appendix A.3 and a union bound over the K columns give $\max_j |A_j| \leq \sqrt{2u} + \frac{2}{3}u\sqrt{h}$ with conditional probability at least $1 - \alpha/2$. Because $hu = o_{\mathbb{P}}(1)$, this bound is $O_{\mathbb{P}}(\sqrt{\log K})$. The studentized bootstrap coordinates are the A_j , recentering only lowers them, and the raw coordinates satisfy $\widehat{\omega}_j A_j \leq \lambda_{\max}(\widehat{\Omega}_\ell)^{1/2}|A_j|$. All three conditional $(1 - \alpha)$ quantiles therefore have positive part $O_{\mathbb{P}}(\sqrt{\log K})$.

Step 3: the stationary bootstrap. Given the data, the stationary bootstrap (Politis and Romano, 1994) draws its restart indicators independently of its start positions. Conditionally on the indicators, the resampled series consists of runs of fixed lengths $r_1, \dots, r_M$ with $\sum_q r_q = n$, whose starts $J_1, \dots, J_M$ are independent and uniform. A run longer than L requires L consecutive positions without a restart, so $\mathbb{P}^*(\max_q r_q > L) \leq n(1 - 1/\ell)^L \leq ne^{-(L-1)/\ell} \leq \alpha/2$ when $L < n$, and the event is empty when $L = n$. On the event $\max_q r_q \leq L$, the bootstrap mean of column j , minus $\bar{Y}_j$ and multiplied by $\sqrt{n}$, equals $n^{-1/2}\sum_q a_j^\top S_{J_q, r_q} = \widehat{\omega}_j U_j$. Conditionally on the lengths, the summands of U_j are independent and centered, because $\sum_s S_{s,r} = 0$, and their total variance satisfies

$$\frac{1}{n} \sum_q \frac{r_q a_j^\top \widehat{\Omega}_{r_q} a_j}{\widehat{\omega}_j^2} \leq \frac{\max_{r \leq L} \lambda_{\max}(\widehat{\Omega}_r)}{\lambda_{\min}(\widehat{\Omega}_\ell)} = 1 + o_{\mathbb{P}}(1).$$

They are bounded in absolute value by $\sqrt{g}$, where $g = \max_{s,r \leq L} \|S_{s,r}\|^2/\{n\lambda_{\min}(\widehat{\Omega}_\ell)\} = O_{\mathbb{P}}[L\{p + \log(nL)\}/n]$. Bernstein’s inequality with $u = \log(4K/\alpha)$ and a union bound over the columns give $\max_j |U_j| = O_{\mathbb{P}}(\sqrt{\log K})$ with conditional failure probability at most $\alpha/2$, because $gu = o_{\mathbb{P}}(1)$. Together with the event for long runs, the conditional failure probability is at most α , and the three conditional quantiles have positive part $O_{\mathbb{P}}(\sqrt{\log K})$ exactly as in Step 2.

Step 4: rejection. The observed raw maximum is at least $\sqrt{n} \bar{f}_W$. The observed studentized maximum is at least $\sqrt{n} \bar{f}_W / \widehat{\omega}_W$, where $\widehat{\omega}_W^2 = 1 + o_{\mathbb{P}}(1)$ by Step 1, which holds uniformly over unit vectors and hence for the data-dependent direction of W . The proof of Theorem 1 gives $\sqrt{n} \bar{f}_W / \sqrt{D} \rightarrow_{\mathbb{P}} c > 0$, so both observed maxima exceed $c\sqrt{D}/2$ with probability tending to one, whereas every critical value above has positive part $O_{\mathbb{P}}(\sqrt{\log K}) = o_{\mathbb{P}}(\sqrt{D})$. Each test therefore rejects with probability tending to one.

A.5 Proof of Proposition 4

Proposition 4 (Complete-log Monte Carlo test). *The test is exact when the log does not depend on the data, yet on our construction it rejects with probability tending to one, even though it uses the true null.*

Formally, (i) for any generator whose log is independent of $X \sim P_0$, $\mathbb{P}_0(p_{\mathcal{L}} \leq \alpha \mid \mathcal{L}) \leq \alpha$ for every $R \geq 0$ and $\alpha \in (0, 1)$; and (ii) under the assumptions of Theorem 1, for fixed $\alpha \in (0, 1)$ and fixed R with $\alpha(R+1) \geq 1$, the test rejects with probability tending to one.

Part (i). The log $\mathcal{L}$ is a function of (X, ω) , and the null draws $X_1, \dots, X_R$ are independent of (X, ω) and of each other. Together with the assumed independence of $\mathcal{L}$ and X , this makes $\mathcal{L}, X, X_1, \dots, X_R$ mutually independent, with each dataset distributed as P_0 . Conditional on $\mathcal{L}$, the map $T_{\mathcal{L}}$ is therefore fixed and the scores $T_{\mathcal{L}}(X), T_{\mathcal{L}}(X_1), \dots, T_{\mathcal{L}}(X_R)$ are i.i.d., hence exchangeable. The rank-counting argument in the proof of Theorem 5 gives $\mathbb{P}_0(p_{\mathcal{L}} \leq \alpha \mid \mathcal{L}) \leq \lfloor \alpha(R+1) \rfloor / (R+1) \leq \alpha$. The conditional bound needs no restriction on the number or form of the logged formulas, and ties only make the test conservative.

Part (ii). Since $\alpha(R+1) \geq 1$, the test rejects whenever $T_{\mathcal{L}}(X) > q := \max_{r \leq R} T_{\mathcal{L}}(X_r)$, because then no null maximum reaches the observed one and $p_{\mathcal{L}} = 1/(R+1) \leq \alpha$. The observed statistic is the maximum studentized score over the complete log, so by Theorem 1 it suffices to show that $q_+ = o_{\mathbb{P}}(\sqrt{D})$.

Condition on (X, ω) , which fixes the logged directions $v_1, \dots, v_K$ with $K \leq (b+1)D$. For a fixed unit vector v_j , the returns $f_{v_j}(t)$ computed from an independent null dataset X_r are independent standard normal variables over t , by (3). Because s_n uses divisor n , $S(X_r, v_j) = \{n/(n-1)\}^{1/2} \tau_{rj}$, where τ_{rj} has a Student t distribution with $\nu = n-1$ degrees of freedom. For $\nu > 4$, $\mathbb{E}\tau_{rj}^4 = 3\nu^2/\{(\nu-2)(\nu-4)\}$, which decreases in ν and equals 25 at $\nu = 5$. Hence, for $n \geq 6$, $\mathbb{E}S(X_r, v_j)^4 \leq (6/5)^2 \cdot 25 = 36$. Markov's inequality and a union bound over the RK pairs (r, j) give, for every $\epsilon > 0$,

$$\mathbb{P}(q > \epsilon\sqrt{D} \mid X, \omega) \leq \frac{36 RK}{\epsilon^4 D^2} \leq \frac{36 R(b+1)}{\epsilon^4 D}.$$

The bound does not depend on (X, ω) and tends to zero as $D \rightarrow \infty$, so $q_+ = o_{\mathbb{P}}(\sqrt{D})$ and the test rejects with probability tending to one. The same bound allows R to grow, provided $R = o(D)$.

A.6 The Deflated Sharpe Ratio

For a reported strategy with per-period Sharpe estimate $\widehat{SR}$ from n observations, the Deflated Sharpe Ratio of Bailey and López de Prado (2014) declares significance at level α when

$$\Phi\left(\frac{(\widehat{SR} - SR_0)\sqrt{n-1}}{\{1 - \widehat{\gamma}_3\widehat{SR} + (\widehat{\gamma}_4 - 1)\widehat{SR}^2/4\}^{1/2}}\right) > 1 - \alpha, \quad SR_0 = \widehat{\sigma}_{SR} a(N),$$

where $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are the sample skewness and kurtosis of the reported strategy's returns, $\widehat{\sigma}_{SR}$ is the standard deviation of the trial Sharpe estimates, N is the number of trials, and

$$a(N) = (1 - \gamma_E)\Phi^{-1}(1 - N^{-1}) + \gamma_E\Phi^{-1}\{1 - (Ne)^{-1}\}$$

with Euler's constant γ_E . When the denominator is not positive, the test does not reject. Applied to our search, the trials are the K logged strategies, the reported strategy is the one with the highest Sharpe estimate, and N is either K or the average-correlation interpolation $\widehat{N} = \bar{r} + (1 - \bar{r})K$, where $\bar{r}$ is the average pairwise sample correlation of the logged return series.

Proposition 6 (Deflated Sharpe Ratio). *Under the assumptions of Theorem 1, suppose that $D \log D = o(n)$ and $0 < \alpha < 1/2$, and put $\sigma_b^2 = 1/\{2(b+1)\} - 4/\{9(b+1)^2\}$. With either choice of N ,*

$$\frac{\sqrt{n} SR_0}{\sqrt{D \log D}} \xrightarrow{\mathbb{P}} \sqrt{2} c \sigma_b > 0,$$

and the test rejects with probability tending to one. The same holds under the planted alternative that adds the mean $\delta = h/\{\sqrt{n}\sqrt{1-\rho}\sqrt{D}\}$ to one coordinate in each batch, for any fixed h .

Proof. Write $Z_j = \sqrt{n} \bar{f}_{v_j}$ for the numerators of the logged scores and $S_j = Z_j/s_n(v_j)$ for the scores. Up to the factor $\{(n-1)/n\}^{1/2}$ that comes from the divisor of the standard deviation, $\sqrt{n}$ times each per-period Sharpe estimate equals S_j ; this factor tends to one and is omitted below.

Step 1: the numerators. The numerators of the basic strategies are $\sqrt{\rho}A_0 + \sqrt{1-\rho}A_i$, and the prefix numerators are $c\sqrt{d} + R_d$ with $R_d = \sqrt{\rho}A_0 + \sqrt{1-\rho}d^{-1/2} \sum_{r \leq d} (H_r - \mu_b)$, where $H_r = \max_{i \in B_r} A_i$ as in the proof of Theorem 1. Because $\sup_d \mathbb{E} R_d^2 < \infty$, Markov's inequality gives $\sum_{d \leq D} R_d = O_{\mathbb{P}}(D)$, $\sum_{d \leq D} \sqrt{d} R_d = O_{\mathbb{P}}(D^{3/2})$, $\sum_{d \leq D} R_d^2 = O_{\mathbb{P}}(D)$, and $\max_{d \leq D} |R_d| = O_{\mathbb{P}}(\sqrt{D})$; the bD numerators of the basic strategies contribute $O_{\mathbb{P}}(D)$ to both the sum and the sum of squares. Since $\sum_{d \leq D} \sqrt{d} = \frac{2}{3}D^{3/2}\{1 + o(1)\}$ and $\sum_{d \leq D} d = \frac{1}{2}D^2\{1 + o(1)\}$, the empirical mean $\bar{Z}$ and variance σ_Z^2 of the $K = (b+1)D$ numerators satisfy

$$\frac{\bar{Z}}{\sqrt{D}} \xrightarrow{\mathbb{P}} \frac{2c}{3(b+1)}, \quad \frac{\sigma_Z^2}{D} \xrightarrow{\mathbb{P}} \frac{c^2}{2(b+1)} - \frac{4c^2}{9(b+1)^2} = c^2\sigma_b^2,$$

and $\max_j |Z_j| = O_{\mathbb{P}}(\sqrt{D})$. The constant σ_b^2 is positive because $9(b+1) > 8$.

Step 2: from numerators to scores. Step 1 of Appendix A.3 gives $\|\hat{\Gamma} - I_p\|_{\text{op}} = O_{\mathbb{P}}(\sqrt{p/n})$, so $\max_j |s_n^2(v_j) - 1| = O_{\mathbb{P}}(\sqrt{D/n})$ and

$$\max_j |S_j - Z_j| \leq \max_j |Z_j| \max_j |s_n(v_j)^{-1} - 1| = O_{\mathbb{P}}(D/\sqrt{n}) = O_{\mathbb{P}}(\sqrt{D}).$$

The empirical mean and standard deviation change by at most the maximum perturbation, up to a factor $\{K/(K-1)\}^{1/2}$, so the standard deviation σ_S of the scores satisfies $\sigma_S/\sqrt{D} \xrightarrow{\mathbb{P}} c\sigma_b$, and the reported score is $\max_j S_j = O_{\mathbb{P}}(\sqrt{D})$. The same operator bound makes every sample correlation between logged returns equal to its population value $\rho + (1-\rho)\langle v_j, v_l \rangle$ up to $o_{\mathbb{P}}(1)$, uniformly over pairs. All directions have nonnegative coordinates, so every population correlation lies in $[\rho, 1]$, and a fraction tending to $b^2/(b+1)^2$ of the pairs are distinct basic strategies with correlation ρ . Hence $\rho - o_{\mathbb{P}}(1) \leq \bar{r} \leq 1 - (1-\rho)b^2/(b+1)^2 + o_{\mathbb{P}}(1)$, so $\hat{N}$ lies between two positive multiples of D with probability tending to one. For both choices of N , the Gaussian quantile asymptotics $\Phi^{-1}(1 - 1/N) \sim (2 \log N)^{1/2}$ give $a(N)/(2 \log D)^{1/2} \xrightarrow{\mathbb{P}} 1$.

Step 3: no rejection. Steps 1 and 2 give $\sqrt{n}SR_0 = \sigma_S a(N)$ and the displayed limit. The reported strategy satisfies $\sqrt{n}\hat{S}R = \max_j S_j = O_{\mathbb{P}}(\sqrt{D}) = o_{\mathbb{P}}(\sqrt{D \log D})$, so $\hat{S}R < SR_0$ with probability tending to one. Because $\alpha < 1/2$, rejection requires a positive numerator in the displayed criterion, which therefore occurs with probability tending to zero.

Step 4: the planted alternative. Adding a constant to a return series changes neither its sample variance nor its sample correlations. Let A_i now denote the coordinate numerators without the planted mean. Selection maximizes $A_i + \sqrt{n}\delta \mathbf{1}\{i \text{ planted}\}$ within each batch, so each selected A_{i_d} lies within $\sqrt{n}\delta$ of H_d , and each round adds a planted mean of at most $\sqrt{n}\delta$ to the prefix. Each prefix numerator therefore differs from the null value $c\sqrt{d} + R_d$ by at most $\sqrt{1-\rho}d^{1/2}\sqrt{n}\delta = h(d/D)^{1/2} \leq h$, and the numerator of each basic strategy changes by at most $h/\sqrt{D}$. Every numerator is therefore perturbed by at most h , which changes $\bar{Z}$, σ_Z , and $\max_j Z_j$ by $O(1)$ and leaves Steps 1–3 unchanged. $\square$

B PROOFS FOR THE METHOD

This appendix states Theorem 5 formally and proves it, extends it to randomization nulls, and describes how the number of null runs affects resolution and power.

B.1 Proof of Theorem 5

Theorem 5 (Validity of GMC). *If every null run repeats the same search procedure, GMC rejects a true null with probability at most α , whatever the procedure and however many null runs are used.*

Formally, let T be the finite real-valued output of a measurable generator that terminates almost surely. Suppose that under P_0 the datasets $X_0, \dots, X_R$ are i.i.d., that $\omega_0, \dots, \omega_R$ are i.i.d. and independent of the data, and that the same map T is applied in every run. Then, for every $R \geq 0$ and $\alpha \in (0, 1)$,

$$\mathbb{P}_0(p_{\mathcal{G}} \leq \alpha) \leq \frac{\lfloor \alpha(R+1) \rfloor}{R+1} \leq \alpha. \quad (10)$$

If ties among the scores occur with probability zero, the first inequality is an equality.

Let $T_r = T(X_r, \omega_r)$ and $c_i = \sum_{j=0}^R \mathbf{1}\{T_j \geq T_i\}$; since $(R+1)p_G = c_0$ is an integer, the test rejects if and only if $c_0 \leq m := \lfloor \alpha(R+1) \rfloor$. The pairs (X_r, ω_r) are i.i.d., so the scores are exchangeable and $\mathbb{P}_0(c_0 \leq m) = (R+1)^{-1} \mathbb{E}_0[\#\{i : c_i \leq m\}]$. At most m indices can satisfy $c_i \leq m$: among $m+1$ such indices, the one with the smallest score would have $c_i \geq m+1$. Without ties, the c_i are a permutation of $1, \dots, R+1$, and exactly m qualify. $\square$

B.2 Validity under Randomization

Let H be a finite group acting measurably on the data space, with $\gamma X \stackrel{d}{=} X$ under the null for every $\gamma \in H$. Set $\gamma_0 = \text{id}$ and draw $\gamma_1, \dots, \gamma_R$ independently and uniformly from H . Apply the same finite-valued, almost surely terminating map to each $\gamma_r X$, with internal randomness $\omega_0, \dots, \omega_R$ that is i.i.d. and independent of X and the transformations. Use the upper-rank p -value (7), which includes the identity run.

To prove validity, let γ_* be uniform on H and independent of everything else, and write $\xi = (\gamma_1, \dots, \gamma_R, \omega_0, \dots, \omega_R)$. Averaging $\gamma X \stackrel{d}{=} X$ over the independent uniform γ_* gives $\gamma_* X \stackrel{d}{=} X$. Because γ_* and X are both independent of ξ , it follows that $(\gamma_* X, \xi) \stackrel{d}{=} (X, \xi)$. Hence the score vector has the same law as

$$(T(\gamma_* X, \omega_0), T(\gamma_1 \gamma_* X, \omega_1), \dots, T(\gamma_R \gamma_* X, \omega_R)).$$

The elements $\gamma_*, \gamma_1 \gamma_*, \dots, \gamma_R \gamma_*$ are independent, uniform on H , and independent of X : the probability that they equal a specified tuple $(a_0, \dots, a_R)$ is $|H|^{-1} \prod_{r=1}^R \mathbb{P}(\gamma_r = a_r a_0^{-1}) = |H|^{-(R+1)}$. Conditional on X , the pairs $(\gamma_r \gamma_*, \omega_r)$ are therefore i.i.d., so the scores are exchangeable conditionally on X , and hence unconditionally. The rank-counting argument in the proof of Theorem 5 then applies verbatim. Duplicate transformations and ties are retained; the weak inequality in (7) makes them conservative. The argument requires joint invariance of the whole data object; symmetry of individual marginal distributions is not enough.

B.3 Number of Null Runs

The smallest possible Monte Carlo p -value is $1/(R+1)$, so a rejection is possible precisely when $R \geq \lceil 1/\alpha \rceil - 1$. For $m = \lfloor \alpha(R+1) \rfloor \geq 1$, the decision is equivalent to $T_0 > \hat{\tau}$, where $\hat{\tau}$ is the m -th largest of the R null scores. Indeed, $T_0 > \hat{\tau}$ holds if and only if at most $m-1$ null scores are at least T_0 , that is, if and only if $(R+1)p_G \leq m$, which is equivalent to $p_G \leq \alpha$ because $(R+1)p_G$ is an integer. For the following calculation, assume the null scores are independent draws from a continuous CDF F_0 and independent of a future target score; randomization scores, which are dependent, are not covered. Then $F_0(T_1), \dots, F_0(T_R)$ are independent uniforms, and because F_0 is nondecreasing, $F_0(\hat{\tau})$ is their m -th largest value. Consequently $V := F_0(\hat{\tau}) \sim \text{Beta}(R+1-m, m)$, and the conditional rejection probability for new null targets is

$$\mathbb{P}_0(T_0 > \hat{\tau} \mid T_1, \dots, T_R) = 1 - V \sim \text{Beta}(m, R+1-m). \quad (11)$$

Its mean, the average conditional size, is $m/(R+1)$ and need not equal α unless the grid aligns; the conditional size given a particular set of null scores need not attain it.

Under an alternative, let the independent target score have CDF F_1 and define $\Psi(u) = \mathbb{P}_{F_1}\{F_0(T_0) \leq u\}$. Since F_0 is nondecreasing, $\{F_0(T_0) > V\} \subseteq \{T_0 > \hat{\tau}\} \subseteq \{F_0(T_0) \geq V\}$, so the two events differ only on $\{F_0(T_0) = V\}$. This event has probability zero because V has a continuous law and is independent of the target. This gives the exact finite- R identity

$$\pi_R = \mathbb{P}_{F_1}(T_0 > \hat{\tau}) = 1 - \mathbb{E}\Psi(V).$$

For completeness, suppose Ψ is twice continuously differentiable on all of $[0, 1]$, with $L_1 = \sup \Psi'$ and $L_2 = \sup |\Psi''|$ finite. Then the oracle power $\pi_\infty = 1 - \Psi(1 - \alpha)$ and the finite- R power π_R satisfy the power-approximation error bound

$$|\pi_\infty - \pi_R| \leq \frac{L_1}{R+1} + \frac{L_2}{8(R+2)} + \frac{L_2}{2(R+1)^2}.$$

The difference $\pi_\infty - \pi_R$ can have either sign, so the bound controls an approximation error, not a loss of power. To see it, set $a = 1 - \alpha$. The beta moments give $0 \leq \mathbb{E}V - a < 1/(R+1)$ and $\text{Var}(V) \leq 1/\{4(R+2)\}$, and Taylor's theorem with Lagrange remainder bounds $|\mathbb{E}\Psi(V) - \Psi(a)|$ by $L_1|\mathbb{E}V - a| + (L_2/2)\{\text{Var}(V) + (\mathbb{E}V - a)^2\}$. The bound depends on the alternative through Ψ and requires global smoothness; it does not show that any particular R , such as $R = 99$, suffices.

C EXPERIMENTAL DETAILS

The simulation, the financial study, and the live-LLM study are distinct experiments, each with a written protocol; the financial study is exploratory. All prespecified endpoints are reported. Exploratory analyses are labeled, and any tests and multiplicity adjustments are stated with them. The tables report counts rather than rounded percentages, so that the Monte Carlo denominator is explicit. Unless stated otherwise, single-rate error bars are two-sided 95% Clopper–Pearson intervals, and intervals for paired differences are descriptive percentile intervals from 20,000 whole-pair bootstrap resamples, not simultaneous confidence statements. Inference across a prespecified family uses the stated test and multiplicity adjustment.

C.1 Simulation Setup

The simulation fixes $n = 2520$, 128 idiosyncratic coordinates, $\rho = 0.75$, $b = 8$, $\alpha = 0.05$, and $R = 19$. Candidate returns and their studentized score are as in (3) and (1); the executable score returns zero if the empirical standard deviation is at most 10^{-12} . The search uses the batches $\{8(d-1)+1, \dots, 8d\}$. It selects each batch’s largest raw mean, with the smaller coordinate breaking a tie, and constructs the normalized equal-weight prefix. Each round logs the evaluations of its eight basic strategies and the prefix. The final report maximizes the studentized score over this complete log, with the earliest log position breaking a tie. The duplicate first prefix is retained. Null-search logs at depths four and eight are the first four or eight rounds of the depth-16 logs, with the retained entries renumbered, so outcomes at different depths are dependent and are not independent replications.

Null A draws every coordinate independently from $N(0, 1)$ on every date. Each calibration dataset is a new full draw. Null B divides the series into 120 independent 21-date blocks. Within a block each coordinate has Gaussian covariance $\Lambda_{rs} = 0.5^{|r-s|}$, and coordinates are independent. A randomization multiplies every coordinate and every date within block j by one common independent Rademacher sign ϵ_j . The finite group of all such block-sign vectors leaves the joint null law unchanged. The observed object uses the identity action; each null run samples a uniform sign vector independently with replacement. Duplicate transformations are kept. These facts meet the conditions of Appendix B.

The holdout first searches 84 complete blocks and then fixes its selected strategy. Its test statistic is $\sqrt{36}$ times the mean of the remaining 36 block sums divided by their unbiased standard deviation, compared with a one-sided t_{35} quantile. Because the blocks are independent in this simulation, the test sample does not feed the search.

The RC-style bootstrap centers each logged return column and draws stationary-bootstrap resamples with restart probability $1/21$: 199 in the size study and 9,999 in the power comparisons (Appendices C.2 and C.3), where recalibration needs a finer p -value grid. Its maximum statistic is the maximum of $\sqrt{n}$ times the resampled column mean. With B resamples, the add-one upper-tail p -value lies on a grid with denominator $B+1$, and the scalar score used for recalibration is $1-p$. This is the stationary bootstrap applied to mean returns in Appendix A.4 with $\ell = 21$; the proposition concerns its ideal conditional quantile, whereas the implementation uses finitely many resamples.

The effective-count rule uses one-sample t marginal p -values and a Šidák adjustment with an effective trial count computed from the correlations. Its empirical correlation eigenvalues are symmetrized and clipped at zero, and its count is restricted to $[1, K]$. Its scalar score is one minus the adjusted minimum p -value. These marginal and numerical conventions differ from the normal-quantile rule of Proposition 2. Because these marginal p -values do not adjust for within-block dependence, its Null B inflation is not an isolated estimate of the effect of adaptivity; Null A and the Gaussian theorem address failure without temporal dependence.

We also report the size of a Giglio–Liao–Xiu-style factor-adjustment/FDR rule (Giglio et al., 2021) and a Harvey–Liu-inspired double-bootstrap rule (Harvey and Liu, 2020), labeled GLX-style and HL-inspired. These labels identify our implemented rules rather than verified reproductions of the complete published procedures.

GMC under Null A. Under Null A, the target and null scores are independent, continuous draws from the same law, and $R = 19$ places $\alpha = 0.05$ exactly on the rank grid, so Theorem 5 gives a rejection probability of exactly 0.05 at every depth. The observed rates of 3.8%, 3.0%, and 3.6% are Monte Carlo estimates of that probability, not the guarantee itself. The depth-8 count, 15/500, lies slightly below 0.05: its two-sided

95% Clopper–Pearson interval is $[0.0169, 0.0490]$, and the unadjusted two-sided exact binomial p -value is 0.0395. Discreteness of the rank grid does not explain this deviation, because the exact rejection probability here is 0.05. The three depths share data and calibration draws, so their counts are not independent. The implementation uses fresh calibration draws, identical scoring in the target and null runs, and exactly the p -value and rejection rule of (7). Three exploratory checks use the same implementation. First, the target’s rank among its 20 scores is, at every depth, consistent with the uniform distribution (chi-square $p = 0.59, 0.50$, and 0.92 ; mean ranks 10.31, 10.38, and 10.33, against 10.5). Second, the 500 replicates have no tied scores, and the target and null scores have the same distribution (Kolmogorov–Smirnov $p = 0.72, 0.90$, and 0.62). Third, 5,000 independent replicates give rejection rates of 5.22%, 4.88%, and 5.02% at depths 4, 8, and 16, with 95% Clopper–Pearson intervals $[4.62, 5.87]$, $[4.30, 5.51]$, and $[4.43, 5.66]\%$, consistent with the exact 5%. The depth-8 count is therefore consistent with a chance deviation.

Table 1: Null-rejection counts at level 5% in the simulation, each out of 500 replicates. Null A is the primary null; Null B adds within-block serial dependence, which the effective-count rule’s marginal tests ignore. The last two columns count any discovery under the complete null; they are not single-winner power endpoints.

Null	Depth	GMC	Holdout	RC-style bootstrap	Effective- count rule	GLX-style	HL-inspired
A	4	19	21	50	158	243	52
A	8	15	19	114	278	350	104
A	16	18	19	264	426	456	241
B	4	23	30	58	331	235	50
B	8	24	30	131	425	363	100
B	16	25	29	293	482	457	257

C.2 Simulation Power Study

The power study uses the Null B design at $D = 16$. For a depth- D alternative, choose one planted coordinate j_d^* uniformly in each searched batch and let $w_D^* = D^{-1/2} \sum_{d=1}^D e_{j_d^*}$. Add the same mean shift

$$\delta(h, D) = \frac{h\sqrt{\Omega_B}}{\sqrt{n}\sqrt{1 - \rho}\sqrt{D}}, \quad \Omega_B = \frac{1}{21} \mathbf{1}^\top \Lambda \mathbf{1} = \frac{59 + 2^{-19}}{21} \approx 2.8095$$

to each planted coordinate. The common factor remains centered, and the oracle direction w_D^* has a signal scale indexed by h . The strength $h = 2.25$ was chosen with the oracle test alone: on the grid $0.5, 0.75, \dots, 3$, it is the strength whose oracle power on 5,000 shared replicates is closest to 0.70, with ties broken toward the smaller h ; the observed oracle power is 0.732. No competing method’s results were used to choose this setting. Under these alternatives, the block-sign randomizations of GMC act on the entire planted dataset, so the null runs in this simulation are not computed on independent unplanted draws.

For each of the three comparators, an independent sample of 4,999 null scores fixes the critical value at its 250th-largest score, and rejection requires strict exceedance. Recalibration thus needs the same kind of null searches that GMC runs, 4,999 instead of 19, so the recalibrated fixed-output rules are benchmarks, not usable post-hoc corrections. Recalibration, the 500 validation replicates, and the 500 paired alternative replicates use different random seeds, so no dataset is used twice. GMC keeps its own rank test. A common 5% level is thus targeted, but neither equal realized null rejection nor equal power is imposed. The recalibrated RC-style bootstrap rejects when its p -value is at most 0.0002 and rejects in 25 of 500 fresh null replicates. The power comparison uses 9,999 resamples instead of the prespecified 199: with 199, the smallest bootstrap p -value occurs in more than 5% of null replicates, so no 5% threshold exists and the rule cannot reject.

The true-discovery endpoint is $\mathbf{1}\{\text{reject}\}\mathbf{1}\{\mu(W) > 0\}$, where $\mu(W)$ is the selected strategy’s population mean in the planted model. It does not require exact support recovery. The primary paired contrast is GMC versus the holdout at $D = 16$, $h = 2.25$. With 55 GMC-only and 34 holdout-only successes, its one-sided exact McNemar p -value is 0.0167; the difference is 4.2 percentage points, with a descriptive paired 95% percentile interval of $[0.6, 7.8]$. The two secondary contrasts form a separate two-test Holm family. GMC versus the recalibrated effective-count rule has adjusted $p = 0.869$. The prespecified contrast with the 199-resample RC-style bootstrap,

adjusted $p = 8.47 \times 10^{-22}$, compares GMC with a rule that cannot reject and is therefore uninformative; against the 9,999-resample version, 26 replicates favor GMC and 21 the bootstrap (one-sided McNemar $p = 0.28$).

Table 2: Power in the simulation study at $D = 16$, $h = 2.25$ after comparator recalibration, each count out of 500 replicates. The null-check column counts rejections on 500 fresh null replicates; for GMC it is the main Null B depth-16 size sample rather than a separate validation sample. The RC-style bootstrap uses 9,999 resamples.

Decision rule	Null check	True discoveries	Power (95% interval)
GMC (own rank test)	25	71	0.142 [0.113,0.176]
Holdout, recalibrated	17	50	0.100 [0.075,0.130]
Effective-count rule, recalibrated	25	78	0.156 [0.125,0.191]
RC-style bootstrap, recalibrated	25	66	0.132 [0.104,0.165]

C.3 Other Signal Strengths and Depths

Table 3 reports the complete signal grid of the simulation rather than only the most favorable point. Replicates are shared across signal strengths, and comparisons away from the primary setting are descriptive. Table 4 changes both the searched depth and the depth-specific planted support and mean shift, so it is not an intervention that changes search depth while holding the alternative fixed. For GMC, the mean fraction of planted coordinates contained in the selected strategy falls from 0.770 at depth four to 0.418 at depth 16, and the mean L^2 norm of its weights on unplanted coordinates grows from approximately 0.355 to 0.756.

Table 3: Exploratory true-discovery counts in the simulation at $D = 16$, out of 500 replicates per row. All non-GMC columns use recalibrated decisions, and the RC-style bootstrap uses 9,999 resamples. The bold row marks the prespecified primary setting.

h	GMC	Holdout	Effective-count rule	RC-style bootstrap
0.50	24	18	24	25
0.75	29	22	32	31
1.00	34	23	34	34
1.25	42	29	38	35
1.50	45	33	47	45
1.75	51	41	52	48
2.00	60	48	64	55
2.25	71	50	78	66
2.50	83	61	89	83
2.75	103	77	112	103
3.00	126	90	134	125

Table 4: Exploratory true-discovery rates in the simulation at $h = 2.25$. Each depth-specific cell has 500 replicates, and comparator scores are recalibrated separately at each depth. The RC-style bootstrap uses 9,999 resamples.

D	GMC	Holdout	Effective-count rule	RC-style bootstrap
4	0.272	0.198	0.330	0.326
8	0.166	0.126	0.194	0.188
16	0.142	0.100	0.156	0.132

C.4 Deflated Sharpe Ratio Simulations

This simulation is exploratory. It uses the simulation design of Appendix C.1 on independent replicates and applies the full DSR of Appendix A.6 at level 0.05, with the complete log as the trials and the highest-scoring strategy as the reported one; N is either the number K of logged trials or the average-correlation interpolation

$\hat{N}$. Here the effective-count rule rejects in 141, 278, and 428 of 500 Null A replicates at depths 4, 8, and 16, in line with the 158, 278, and 426 of Table 1.

Table 5 reports the DSR’s null rejections under Null A. They increase with depth for both choices of N , from near 5% at $D = 4$ to more than 90% at $D = 1024$, because the dispersion term raises the threshold more slowly than the drift raises the reported score. By the leading-order calculation behind Proposition 6, the threshold catches up with the drift only near $D \approx 1.6 \times 10^4$ with $N = K$ and $D \approx 6.4 \times 10^4$ with $\hat{N}$.

Table 5: Exploratory: null rejections of the full Deflated Sharpe Ratio under Null A, out of 500 replicates per depth, at level 0.05. N is the number of trials in the DSR threshold.

D	4	8	16	32	64	128	256	512	1024
$N = K$	26	36	68	112	217	311	389	448	465
$N = \hat{N}$	53	80	137	210	328	441	486	499	500

For power, the prespecified setting of Appendix C.2 (Null B, $D = 16$, $h = 2.25$) is used on independent replicates. At level 0.05, the DSR rejects in 165 ($N = K$) and 245 ($\hat{N}$) of 500 fresh null replicates. Recalibrated as in Appendix C.2, with the 250th largest of 4,999 null statistics as its critical value, it rejects in 30 and 29 of 500 null replicates and makes 51 and 59 true discoveries in 500 alternative replicates. On the same replicates, GMC rejects in 29 null replicates and makes 80 true discoveries, in line with the 25 and 71 of Table 2. These counts are descriptive, and no paired test is reported.

C.5 Holdout Split Points

This check is exploratory. It uses the prespecified power setting of Appendix C.2 (Null B, $D = 16$, $h = 2.25$) on 1,000 independent replicates and varies the holdout’s split. The holdout searches the first m of the 120 blocks, reports the highest-scoring log entry there, and applies the one-sided t -test to that strategy’s block sums on the remaining blocks; because the blocks are independent, this test is exact, so no recalibration is used. GMC runs on the same replicates and makes 145 true discoveries. Table 6 reports the holdout’s true discoveries, the discordant pairs, and one-sided exact McNemar p -values for GMC over the holdout, unadjusted for the five comparisons. GMC makes at least as many true discoveries at every split point, but against the even split its advantage is about one percentage point and is not significant. At the split of the main study ($m = 84$), the difference is 1.7 points ($p = 0.095$), smaller than the 4.2 points of Table 2, where the recalibrated holdout rejects in 17 of 500 null validation replicates, below 5%.

Table 6: Exploratory: true discoveries out of 1,000 replicates at the prespecified power setting, by holdout split. GMC makes 145 on the same replicates. Discordant pairs count replicates in which only one method made a true discovery; p is the one-sided exact McNemar p -value for GMC over the holdout.

Search blocks m	Test blocks	Holdout	GMC only	Holdout only	p
36	84	123	74	52	0.030
60	60	134	78	67	0.20
84	36	128	83	66	0.095
96	24	110	102	67	0.004
108	12	80	120	55	5.0×10^{-7}

C.6 Financial Study

The financial study follows a separate prespecified protocol and is exploratory relative to the paper’s primary analyses. It asks how GMC and a holdout compare when a deterministic feedback-driven miner searches real financial features. A *strict discovery* requires rejection and a selected signal whose direction is close to the planted one (cosine similarity at least 0.90); the selected formula need not match the planted formula.

Results. The results show a crossover rather than uniform improvement (Figure 4). For the momentum-aligned signal, GMC strict-discovery rates are 10.8%, 36.2%, and 73.4% across the three strengths, whereas the

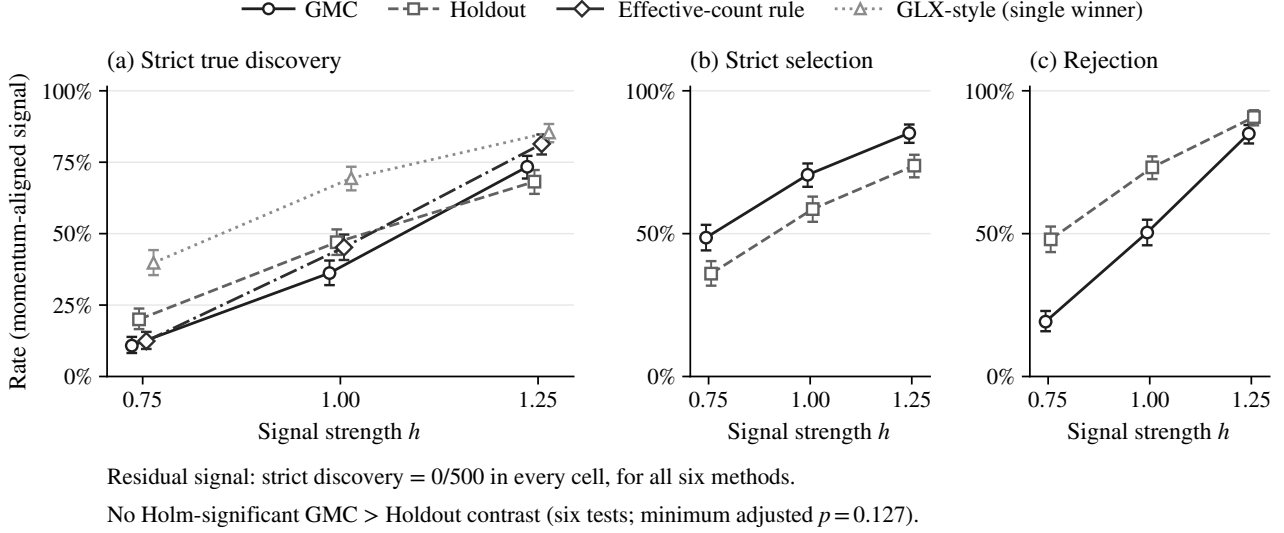


Figure 4: Financial study: better selection need not produce more strict discoveries. Left: strict-discovery rates for the momentum-aligned signal; the recalibrated RC-style bootstrap and HL-inspired rule are zero throughout. Middle and right: GMC selects the aligned direction more often than the holdout but rejects less often at weaker signals, and a strict discovery requires both. Exploratory; 500 replicates per cell; 95% Clopper–Pearson intervals.

holdout reaches 20.0%, 47.0%, and 68.2%. The high-signal difference has an unadjusted one-sided $p = 0.0211$ but a Holm-adjusted value of 0.127, and none of the six prespecified one-sided GMC-versus-holdout tests rejects after adjustment. For the residual signal, every method makes zero strict discoveries at every tested strength. Separating selection from rejection clarifies the pattern for the aligned signal. At the middle strength, GMC selects a direction meeting the cosine criterion in 70.6% of replicates, compared with 58.6% for the holdout, yet their rejection rates are 50.4% and 73.2%. The higher selection rate of GMC is thus offset by its lower rejection rate in the joint endpoint. With $R = 99$ null runs, GMC makes 37.0% strict discoveries in 200 fresh replicates, versus 36.2% at $R = 19$ (Appendix C.7). The remainder of this subsection specifies the protocol.

Protocol. The data originate from Yahoo Finance (<https://finance.yahoo.com/>). The fixed panel has 1,328,003 eligible rows, 30 features, and 2,765 scoring dates from January 2, 2015 through December 30, 2025. Securities come from an acquisition-date static S&P 500 roster; the panel is not reconstructed with point-in-time membership. The study is therefore not a survivorship-free, cost-adjusted trading backtest, and its guarantees concern the simulation law below, not actual market returns.

For each simulated null replicate, independently and uniformly permute finite next-day return labels among eligible securities within each date. Keep date groups, features, return multisets, and row order fixed. The target and null datasets are drawn independently from this law. Under an alternative, add $\delta_{\text{plant}}(h)z_{\text{plant}}$ to the permuted return vector, where z_{plant} is a fixed planted signal. The aligned signal is within-date standardized 63-day momentum. The residual signal starts from standardized 21-day maximum return, subtracts its no-intercept least-squares projection on standardized momentum at 21, 63, and 126 days, one-day reversal, 21- and 63-day volatility, and 21-day log dollar volume, and is restandardized within date. The Moore–Penrose inverse uses cutoff 10^{-12} . Coefficients and the row-level signal are fixed before simulation.

Signal amplitude is $\delta_{\text{plant}}(h) = h \text{se}_{\text{plant}}/a_{\text{plant}}$, with a_{plant} the mean response of the planted signal’s portfolio to its own signal. The scale se_{plant} is the Euclidean norm of chronological 21-day block sums of the unpermuted portfolio returns, divided by 2,765. Portfolio weights are within-date demeaned and standardized, then normalized to unit L^1 norm. The cosine criterion below compares row-level signal vectors, not these portfolios.

The deterministic miner uses a fixed list of 640 basic strategies, each defined by one expression, in 16 rounds of 40. The expression grammar includes 30 raw features, sign, hyperbolic tangent, cross-sectional rank and standardization, rolling mean/sum/standard deviation at windows 5, 21, and 63, and pairwise sum/difference/product. Raw features lead the list. The other expressions are written in a standard form and sorted by a hash computed with

the fixed public seed 202607290401, and the first 610 complete the list. Round winners feed prefix combinations. The complete log consists of all 640 basic strategies and 16 prefixes, and the highest-score entry is reported, with earliest-position tie-breaking. This search uses feedback but has no live LLM calls.

The score is the sum of chronological 21-day block return sums divided by their Euclidean norm. The holdout searches the first 91 complete blocks (1,911 dates) and evaluates on the remaining blocks, including the final short block. GMC uses all dates and, under alternatives, 19 null runs on independent *unplanted* null datasets. This differs from the simulation, whose null runs act on sign transformations of the planted dataset. Five non-GMC scores are recalibrated on 4,999 null replicates, and all six decisions are checked on 500 fresh validation replicates. The two additional fixed-output rules, adapted to report a single strategy, are a GLX-style maximum residualized statistic and an HL-inspired minimum double-bootstrap candidate p -value. They are not complete replications of the published discovery-set methods.

The strengths were chosen with the holdout alone, on 100 replicates for each of two planted signals and 13 strengths on the grid 0.5, 0.75, $\dots$, 3, 3.5, 4. The strength whose strict-discovery rate is closest to 0.50 determines three consecutive strengths on the grid, centered on it; ties go to the lower strength, and at an edge of the grid the first or last three strengths are used. The resulting strengths are 0.75, 1, 1.25 for the aligned signal and 0.5, 0.75, 1 for the residual signal. The reported replicates use datasets generated from random seeds not used for this choice, for recalibration, or for validation. The six cells, six methods, and 500 replicates per cell yield 18,000 method-level rows.

For each candidate, the row-level direction is its within-date demeaned signal divided by the within-date population standard deviation. Ineligible entries and zero-variance dates are filled with zero. A prefix direction is the sum of its component directions divided by the square root of their count. Let d_W and z_{plant} be the selected and planted row-level vectors restricted to rows where the planted signal is defined and both vectors are finite. Zero-filled candidate entries remain in this comparison, which is not further restricted to rows where the candidate itself is eligible. The selected direction meets the cosine criterion, a *strict selection*, when

$$\frac{\langle d_W, z_{\text{plant}} \rangle}{\|d_W\|_2 \|z_{\text{plant}}\|_2} \geq 0.90.$$

An empty comparison set or a zero norm receives cosine zero. The sign is not reversed to favor the planted direction, and neither an absolute cosine nor an expression-name match is used. A strict discovery requires both a strict selection and rejection; a method can reject without a strict selection, or make one without rejecting. For the residual signal, GMC’s selected direction has a positive cosine with the planted signal in 74.2–89.8% of replicates but never reaches 0.90, so a positive cosine cannot stand in for the criterion.

Additional comparators. Both rules operate on the current log’s 656-column return matrix, aggregated into 132 chronological blocks (the last has 14 dates). Neither regenerates formulas inside its own bootstrap or factor adjustment. For the GLX-style rule, center each column of the block matrix Z and subtract its rank-one singular-value decomposition fit. With residual sample standard deviation s_j^{res} using divisor 131, form $t_j = \sqrt{132} \bar{Z}_j / s_j^{\text{res}}$. Only finite columns with $s_j^{\text{res}} > 10^{-12}$ are active. The single-winner score is the largest active t_j , with earliest-log tie-breaking. Its nominal discovery set uses the first grid threshold $\theta = 0, 0.01, 0.02, \dots$ up to the grid ceiling of $\sqrt{2 \log 656}$ at which

$$\widehat{\text{FDP}}(\theta) = \min \left\{ 1, \frac{656 \bar{\Phi}(\theta)}{\max\{|\{j : j \text{ active, } t_j \geq \theta\}|, 1\}} \right\} \leq 0.05,$$

or the empty set if there is no such threshold. The factor rank is fixed at one, not estimated.

The HL-inspired rule instead uses ordinary column sample standard deviations and 199 outer by 199 inner row-bootstrap resamples. Outer samples resample the centered block matrix and give maxima $M_b = \max(0, \max_j t_{bj}^*)$. Recenter each outer sample and resample it again to obtain $M_{b\ell}^{**}$. The two-stage values are

$$q_b = \frac{1 + \sum_{\ell=1}^{199} \mathbf{1}\{M_{b\ell}^{**} \geq M_b\}}{200}, \quad p_j = \frac{1 + \sum_{b=1}^{199} \mathbf{1}\{M_b \geq t_j\}}{200}, \quad \tilde{p}_j = \frac{1 + \sum_{b=1}^{199} \mathbf{1}\{q_b \leq p_j\}}{200}.$$

Inactive columns receive $\tilde{p}_j = 1$. The single-winner score is $1 - \min_j \tilde{p}_j$, again with earliest-log ties; the nominal discovery set contains active columns with $\tilde{p}_j \leq 0.05$. The recalibration of these *single-winner scores* at a common level is separate from their nominal discovery sets. A nominal set that contains any truly aligned entry is not counted as a rejection with a true reported entry.

Table 7: All strict-discovery counts in the financial study, out of 500 replicates. Every comparator is recalibrated on null replicates. Each zero cell has a two-sided 95% Clopper–Pearson upper bound of 0.0074.

Method	Momentum aligned			Residual maximum return		
	$h = 0.75$	1.00	1.25	0.50	0.75	1.00
GMC	54	181	367	0	0	0
Holdout	100	235	341	0	0	0
Effective-count rule	62	226	407	0	0	0
GLX-style (single winner)	199	347	427	0	0	0
RC-style bootstrap	0	0	0	0	0	0
HL-inspired (single winner)	0	0	0	0	0	0

Table 8: The complete prespecified six-test family, all one-sided for GMC superiority over the holdout. Effect intervals are descriptive paired 95% percentile intervals, not Holm-adjusted intervals. A degenerate interval at zero with no discordant observations is not an equivalence result.

Cell	Difference	Interval	GMC-only / holdout-only	Raw p	Holm p
Aligned 0.75	−0.092	[−0.130, −0.054]	25 / 71	> 0.999	1
Aligned 1.00	−0.108	[−0.156, −0.060]	52 / 106	> 0.999	1
Aligned 1.25	0.052	[0.004, 0.100]	89 / 63	0.0211	0.127
Residual 0.50	0	[0, 0]	0 / 0	1	1
Residual 0.75	0	[0, 0]	0 / 0	1	1
Residual 1.00	0	[0, 0]	0 / 0	1	1

Null validation. On fresh null replicates, rejection counts are 26 for GMC, 36 for the holdout, 21 for the effective-count rule, 25 for the GLX-style rule, and zero for both bootstrap-based rules, each out of 500: with 199 resamples, their p -values take the smallest grid value in 90% (RC-style) and 19% (HL-inspired) of calibration replicates, so their recalibrated thresholds cannot be exceeded. The realized rate of the holdout, 7.2%, is higher than the 5.2% of GMC; the comparison targets a common 5% level, not equal realized sizes, and this difference may contribute to the higher rejection frequency of the holdout. A prespecified screen for excess null rejection, Bonferroni-adjusted over six tests, flags 38 or more rejections for GMC (binomial reference) and 39 or more for the recalibrated methods (beta-binomial reference with parameters (500, 250, 4750)). No method is flagged; passing this screen does not certify an exact 5% conditional size. Before recalibration, the null-rejection rates at level 5% of the effective-count rule, RC-style bootstrap, GLX-style rule, and HL-inspired rule are 99.2%, 100%, 98.8%, and 72.0%; Table 7 reports the recalibrated rules.

Selection versus rejection. For the three aligned cells, GMC strict-selection counts are 243, 353, and 426, and the holdout’s counts are 180, 293, and 369. Their rejection counts are 96, 252, and 425 versus 240, 366, and 454. These marginal counts describe the trade-off in the joint endpoint without establishing which isolated mechanism causes it. Other contrasts between methods, including the favorable GLX-style numbers, remain descriptive.

C.7 Sensitivity to R

Two separate checks use $R = 99$. Under Null B of the simulation at depth 16, GMC rejects in 15 of 200 replicates, or 7.5% with interval [4.26%, 12.07%]; this is a small null-size sensitivity sample. In the financial study, 200 fresh alternative replicates per planted signal give 74 strict discoveries for the aligned signal at $h = 1$, or 37.0% with interval [30.30%, 44.09%], and zero for the residual signal at $h = 0.75$, with interval [0, 1.83%]. For the aligned signal, rejection and strict selection occur in 101 and 138 of 200 replicates; for the residual signal, rejection occurs in 10 of 200 and strict selection never occurs. These separate samples establish neither equivalence between $R = 19$ and $R = 99$ nor a general power gain from larger R .

A supplementary simulation of reused null runs draws 20 independent sets of $R = 19$ null runs and evaluates 200 new Null A targets against each set. The rejection rates of the sets range from zero to 0.14, with median

0.05 and sample standard deviation 0.040. This illustrates the conditional variability in (11); it neither refutes the guarantee of Theorem 5, which averages over the null runs, nor validates an arbitrary fixed set.

C.8 Live LLM Study

Protocol. The live study follows a written protocol that specifies the model, prompts, data, sample sizes, endpoints, tests, and interval rules. The signal strengths $h \in \{4, 8\}$, the definition of a true discovery, and the interval rule for the power study were set using pilot searches that are not in the reported sample.

Model and generation. The model is DeepSeek-V4.1-Flash, requested as `deepseek-flash` through DeepSeek’s API (<https://www.deepseek.com/>) with thinking disabled, no reasoning-effort field, an 8,192-token output limit, and JSON-object output. Every accepted reply carried the pinned system fingerprint `aeb56401ca74e127821c4f9126dcb669`. Each search has eight rounds, and each round sends one request for 40 formulas in a fixed expression language over 30 features, with arithmetic, elementwise functions (absolute value, exponential, logarithm, sign, square root, and clipping), cross-sectional rank and z -score, lags, and rolling means, sums, standard deviations, maxima, and minima. A proposal fills a valid slot if it parses, is not a duplicate within its reply, and meets fixed complexity limits: at most 10 syntax-tree nodes, depth four, two time-series and two cross-sectional transforms, and rolling windows summing to at most 42 days. Per-reply limits apply only to time-series operators, which bound computation. The system prompt is identical with and without feedback. With feedback on, a round’s prompt lists the 20 distinct valid formulas with the highest scores so far, in descending order with scores to three decimals. With feedback off, it lists the 20 most recently introduced distinct valid formulas in generation order, without scores, ranks, or winners. Both conditions also report the previous round’s counts of invalid, duplicate, and missing slots, which depend only on syntax and the features, and a random token unique to each search. There is no fallback formula: invalid slots stay invalid. An unusable reply, one that is truncated or is not a JSON object with a candidate list, is retried at most twice more in the same round, after which the round’s 40 slots are invalid. Every search starts from a fresh context, and no information passes between searches.

Data, null, and score. The features come from the Yahoo-derived panel of Appendix C.6; the returns are simulated. Formulas are evaluated on the full 2,765-date panel, so that rolling operators have warm-up data, but only the last 2,520 scoring dates, with 455–499 securities per date, form positions and returns. On each date, a formula’s finite values are standardized cross-sectionally and divided by the sum of their absolute values, which gives a market-neutral portfolio with unit gross exposure; dates with fewer than two finite values or zero cross-sectional variance receive no position. Under the null, returns are independent standard normal variables for every security and date, independent of the features, so every formula chosen without looking at the returns has zero expected return. A formula’s score is the studentized mean (1) of its daily portfolio returns, recomputed with its own denominator on every dataset. A search reports the maximum score over its complete log, with ties broken by first appearance; a formula with no position on any date cannot win and receives the lowest score.

Comparators. For each null search, the complete-log Monte Carlo test re-scores every distinct valid formula in the log on 19 fresh null return panels and applies (6). The panels are generated date by date from seeds that depend only on the dataset and the date. The searches with and without feedback on the same dataset therefore share their panels, different datasets never do, and the first 19 of the 999 panels used in the sensitivity analysis are the primary ones. The effective-count rule and the RC-style bootstrap are those of Appendix C.1, applied to the daily returns of the distinct valid formulas in the log; the RC-style bootstrap uses 999 stationary-bootstrap resamples. For the curve across search rounds in Figure 3, we recompute the complete-log Monte Carlo test using only the first d rounds of each log, for the target and the null panels alike.

Tests and null check. The primary family has two one-sided tests at family level 0.05 with Holm adjustment: an exact binomial test that the null rejection rate with feedback exceeds 0.05, and an exact McNemar test that paired searches reject more often with feedback than without. The interval for the paired difference combines two-sided 97.5% Clopper–Pearson intervals for the two discordant probabilities. The GMC null check divides the 1,000 feedback searches at random into 50 disjoint groups of 20, each with a designated target. An alarm requires at least 7 rejections. The 50 groups are a seeded random permutation of the searches, so each designated target is a uniformly random member of its group and, absent ties, the number of rejections is $\text{Binomial}(50, 1/20)$ whatever the model does; an alarm has probability 0.012. The check therefore cannot detect a difference between target

and null searches and is not a test of exchangeability; the check of GMC’s false-positive rate below addresses that question.

Power study. Let w_g be the position matrix of the planted formula

$$(\text{cs_zscore}(\text{mom21}) \text{ sign}(\text{rev1})) - 0.5 \text{ cs_rank}(\text{vol5}).$$

Returns at strength h are $\mu_h + \varepsilon$ with $\mu_h = h w_g / \|w_g\|_F$ and null noise ε , so the known-variance z -statistic of the oracle portfolio has noncentrality exactly h . For a selected portfolio w_f , the planted effect is $\lambda = \langle w_f, \mu_h \rangle / \|w_f\|_F = h \cos(w_f, w_g)$, computed over the full window for both methods. The primary endpoint is a true discovery: rejection with $\cos(w_f, w_g) \geq 0.25$, the same similarity criterion at both strengths; the secondary endpoint is rejection with $\lambda \geq 0.5$. The holdout runs the same feedback search using only the first 1,260 dates and rejects if the selected portfolio’s known-variance z -statistic on the last 1,260 dates exceeds $\Phi^{-1}(0.95)$. Each target search and its holdout search share the same simulated returns.

Estimating GMC power. Fresh null runs for each of the 600 signal targets, 19 per target, would require 11,400 searches. The protocol instead estimates GMC’s power, averaged over the randomness of the null runs, from the pool of $M = 1,000$ null searches with feedback. If $k(t)$ pool scores lie strictly below a target score t , a uniformly drawn 19-member subset of the pool rejects with probability $q_{19}(t) = \binom{k(t)}{19} / \binom{M}{19}$, and the expectation of $q_{19}(t)$ over the pool equals the rejection probability with 19 independent fresh null runs. The estimate at each strength is $N^{-1} \sum_{i=1}^N q_{19}(t_i) \mathbf{1}\{\cos(w_{f_i}, w_g) \geq 0.25\}$ over its $N = 300$ targets; it is an average probability, not a count of rejections, and $R = 99$ is a sensitivity analysis. Intervals come from a two-sample bootstrap with 10,000 resamples that draws pool members and paired targets jointly and recomputes every q . Because this bootstrap undercovers when discoveries are rare, the protocol sets a rule for the interval. The paired bootstrap is used when both methods have at least 10 true discoveries, counting the holdout’s discoveries and GMC’s expected number, N times its estimated power. Otherwise, a 98.75% Clopper–Pearson interval for the holdout is combined, by Bonferroni, with a 98.75% interval for GMC: its bootstrap interval if GMC has at least 10 expected discoveries, and if not, the interval from zero to an upper Clopper–Pearson bound on the probability that a target search selects a formula meeting the similarity criterion. Every true discovery requires such a selection, so this probability caps GMC’s power. In simulations covering six scenarios, the 97.5% intervals for the power difference at the two strengths both covered their true differences in 94.9% to 100% of replicates. The rule is a prespecified approximation, not an exact interval. Both primary contrasts meet the condition for the paired bootstrap, with 10 and 43 holdout discoveries and 11.9 and 39.8 expected GMC discoveries.

Table 9: Live study: null rejections out of 1,000 searches for each rule, with and without feedback, and the prespecified tests. The GMC check has denominator 50, and its count without feedback is exploratory.

Rule	Without feedback	With feedback	Role
Complete-log Monte Carlo, $R = 19$	49	196	primary
Complete-log Monte Carlo, $R = 999$	48	234	sensitivity
Effective-count rule	268	620	descriptive
RC-style bootstrap	59	73	descriptive
GMC null check (of 50)	3	3	alarm at ≥ 7
Exact binomial test, rate with feedback > 0.05 : $p = 3.9 \times 10^{-60}$ (Holm 7.7×10^{-60}).			
Exact McNemar test, with feedback $>$ without (170 versus 23 discordant pairs): $p = 3.4 \times 10^{-29}$ (Holm 3.4×10^{-29}).			
Paired difference 0.147; conservative 95% interval [0.108, 0.185].			

Other results. After rounds one to eight, the complete-log Monte Carlo test, computed using only the rounds completed so far, gives rejection counts of 50, 71, 102, 125, 136, 159, 182, and 196 with feedback, and 51, 51, 48, 51, 46, 47, and 49 without. Exploratory score summaries, means over the 1,000 searches of each kind, distinguish the target from its fixed-output reference distribution. Without feedback, the average target maximum and the average maximum from re-scoring its formulas on fresh null data were both 2.67. With feedback, the former rose to 2.91 while the latter fell to 2.45, consistent with the model concentrating on fewer

Table 10: Live study: power at the two planted strengths, in percent. GMC rates are pooled estimates of average power. Intervals are 95% bootstrap percentile intervals, except that the two primary differences carry the prespecified 97.5% intervals.

Endpoint	$h = 4$			$h = 8$		
	GMC	Holdout	Difference	GMC	Holdout	Difference
True discovery (primary)	4.0	3.3	0.6 [−2.6, 3.9]	13.3	14.3	−1.1 [−7.3, 5.3]
Rejection with $\lambda \geq 0.5$	4.3	3.3	0.9 [−1.8, 3.8]	20.2	19.3	0.9 [−5.3, 7.0]
Rejection	10.4	10.7	−0.3 [−4.6, 4.1]	22.9	22.0	0.9 [−5.6, 7.4]
Selects $\cos \geq 0.25$	8.0	5.0		15.0	15.3	
True discovery, $R = 99$	4.1	3.3	0.8 [−2.1, 3.7]	13.5	14.3	−0.8 [−6.4, 4.7]

distinct formulas (192 versus 264 per search). These re-scored maxima hold the target’s formulas fixed; they are not winners from new GMC searches. With feedback, 55% of reported winners came from the last three rounds, compared with 39% without, consistent with feedback steering later proposals toward high in-sample scores.

Audit and cost. The 3,200 searches of the feedback and power comparisons made 25,690 API attempts: 25,599 accepted replies, 90 unusable replies that were retried, and one transient network failure that was retried. One round of one search received no usable reply in three attempts, so its 40 slots are invalid. Every accepted reply reported the pinned model and fingerprint. Of the 1,024,000 formula slots, 944,400 (92.2%) held valid formulas: 91.5% in the null searches with feedback, 93.4% without feedback, and 91.9% in the power study. Of the remaining slots, 77,844 held proposals that failed to parse, used an unknown feature, or exceeded a per-formula complexity limit, and 1,756 were missing, duplicated, malformed, or over a per-reply limit. These searches used 27.8 million input tokens, of which 9.9 million were cache hits, and 16.9 million output tokens, for US\$20.51 in API fees. A search took a median of 4.8 minutes on one CPU worker, so a GMC decision with $R = 19$, which requires 20 searches, costs about 1.6 worker-hours and US\$0.13. All 3,200 planned searches completed. These totals exclude the two separately generated sets of 1,000 null searches used below to estimate GMC’s false-positive rate.

GMC false-positive rate. The null check above cannot detect a difference between target and null searches, and using the 1,000 searches as both targets and null runs gives a pooled rejection rate of exactly $1/20$ by algebra. An informative check needs null runs generated separately from the target searches, which is also how GMC would be used. We therefore use, for each feedback policy, a separately generated set of 1,000 null searches on independent null datasets, with the same code, prompts, and pinned model version; they omit only the re-scoring and comparator computations, which do not feed back into a search. The two sets share their null datasets, as the two policies do in Table 9, so the two estimates are not independent. The estimator, interval, and reading of the result were fixed in advance. For each policy, the estimand is the probability that GMC with $R = 19$ rejects when the target is one of that policy’s 1,000 null searches reported in Table 9 and the null runs come from its separate set. As in the power study, it is estimated by averaging, over the 1,000 target searches, the probability that the target exceeds every member of a uniformly drawn 19-member subset of the separate set, with the same two-sample bootstrap.

The estimate is 4.52% without feedback (bootstrap standard error 0.64 percentage points; approximate 95% interval 3.37–5.86%) and 4.56% with feedback (0.65; 3.43–5.95%), both consistent with the nominal 5%; Table 11 adds the sensitivity analyses. None of the four sets has tied scores or a search without a valid formula, and under each policy the score distributions of the two sets are close (without feedback, Kolmogorov–Smirnov $p = 0.95$ and means 2.67 and 2.68; with feedback, $p = 0.69$ and means 2.91 and 2.93). Every accepted reply in the separate sets reported the pinned model and fingerprint, and 93.1% and 92.3% of their formula slots held valid formulas.

These results are empirical checks of the exchangeability condition in one configuration, not a proof of error control. In simulations with Gaussian scores, such an interval excludes 5% with probability 0.30 when the true rate is 6%, 0.78 at 7.2%, and 0.93 at 7.8%, so an inflation of one percentage point would usually go undetected; its simulated coverage is 0.94–0.95. The statements are limited to one model version and simulated Gaussian returns on real features.

Table 11: Estimated GMC null rejection rates in percent for the separately generated null searches, with bootstrap standard errors and approximate 95% intervals. Under each policy, only the first row enters the prespecified reading of the result.

Targets and null runs	Estimate	SE	Interval
<i>Without feedback</i>			
Targets from Table 9, null runs from the separate set, $R = 19$	4.52	0.64	[3.37, 5.86]
Targets from Table 9, null runs from the separate set, $R = 99$	4.15	0.74	[2.84, 5.74]
Targets from the separate set, null runs from Table 9, $R = 19$	5.46	0.74	[4.13, 7.04]
Targets from the separate set, null runs from Table 9, $R = 99$	5.91	0.97	[4.14, 7.94]
<i>With feedback</i>			
Targets from Table 9, null runs from the separate set, $R = 19$	4.56	0.65	[3.43, 5.95]
Targets from Table 9, null runs from the separate set, $R = 99$	4.31	0.78	[2.96, 6.00]
Targets from the separate set, null runs from Table 9, $R = 19$	5.39	0.72	[4.10, 6.89]
Targets from the separate set, null runs from Table 9, $R = 99$	5.65	0.91	[4.02, 7.53]

C.9 Compute and Reproducibility

All experiments run on CPUs; live model inference is delegated to the provider API. The API cost of the live-LLM study is reported in Appendix C.8. Isolated compute measurements for the simulation use 20 workers, CPU affinity 0–23, chunks of 50 replicates, and process-tree user-plus-system CPU time. Each method runs without shared search-generation caches. These measurements account for workloads; they are not speed comparisons on identical workloads, and the hardware model and a complete environment lockfile are not reported.

Table 12: Isolated CPU accounting for the simulation. Workload differences are part of the reported comparison. These numbers are not the runtimes of the live-LLM or financial-study runs. The RC-style row uses 199 resamples.

Method	CPU-hours	Workload
GMC	6.83	500 null-size and 500 power replicates, including null runs
Holdout	1.78	4,999 recalibration, 500 validation, and 500 power replicates
Effective-count rule	2.13	4,999 recalibration, 500 validation, and 500 power replicates
RC-style bootstrap	2.33	4,999 recalibration, 500 validation, and 500 power replicates

Measured in complete search runs, serial calibration with fresh null runs makes $R + 1$ calls to the search. If every run takes time at most $C_{\mathcal{G}}$ and memory at most $\mathcal{M}_{\mathcal{G}}$, calibration uses $O((R + 1)C_{\mathcal{G}})$ time and $O(\mathcal{M}_{\mathcal{G}} + R)$ memory when scores are retained. For a fixed- R rejection alone, null scores can instead be counted online, reducing additional score storage to $O(1)$. Parallel execution changes elapsed time and peak memory, not the required number of search executions. A random stopping time is allowed only when each run terminates almost surely and follows the same rule; Theorem 5 does not bound the expected runtime.

Our reimplementations of the Nikolopoulos (2026) workflow is not part of the prespecified comparisons. It reproduces 23 of the 25 published reference values we checked, so its fidelity is not established. After recalibration, it rejects in 32 of 500 fresh null replicates and makes 4 true discoveries in 500 alternative replicates, compared with 71 for GMC in the simulation. These observations are exploratory and come from our reimplementations of the published description, not from a validated run of the original authors’ method.

The Gaussian simulation can be regenerated from its specified law. The financial panel and provider outputs are not redistributed, because their redistribution requires a separate licensing review.