How Many Strategies Did You Really Try? Effective Trial Counts for LLM-Proposed Backtests

Working draft. Companion code: `research/effective_trials.py`*. Sequel to the ÀYÒ out-of-sample gating preprint.*

Abstract

Selection-bias corrections for backtested trading strategies, from the Deflated Sharpe Ratio to simple trials-adjusted performance bars, all take one key input: the number of trials `N`. When a large language model is the strategy-search process, `N` is neither known nor safely guessable. The model emits many candidate strategies, but they are highly correlated, so the raw proposal count over-states how many independent bets were actually placed. I import the *effective number of tests* estimator from statistical genetics, computed from the eigenvalues of the candidates' return-correlation matrix, as a principled effective trial count `M_eff`. I show where this correction matters and where it does not: it is first-order in the log-style trials-adjusted bars practitioners actually deploy, and second-order in the Deflated Sharpe Ratio, whose benchmark grows only as the square root of the log of `N`. My central empirical finding, on the real ÀYÒ pipeline, is that the size of the correction is a property of the *generation protocol*: a diverse prompting protocol yields near-independent proposals (`M_eff` approximately `m`, no correction needed), while drawing repeatedly from a single narrow angle collapses fourteen proposals to roughly three effective trials, inflating the selection bar by about 0.33 Sharpe and flipping real deployment decisions. `M_eff` is the instrument that tells a practitioner which regime their generator is in.

---

1. Introduction

The failure mode of amateur quantitative research is well known: try many strategies on one dataset, keep the best, and mistake the reward-for-searching for a genuine edge. The standard defenses raise the bar in proportion to how many strategies were tried. Bailey and López de Prado's Deflated Sharpe Ratio (DSR) deflates the best observed Sharpe by the expected maximum of `N` trial Sharpes. Practitioner gates, including the one in ÀYÒ, raise a required out-of-sample Sharpe as a function of `N`.

All of these assume `N` is known. That assumption quietly breaks when the search process is a language model. An LLM asked for trading strategies does not draw `N` independent hypotheses. It produces many variations on a few themes, momentum here, mean reversion there, each re-dressed. The raw proposal count `m` therefore over-states the independent search effort, and any correction keyed to `m` is

miscalibrated. This paper asks: what is the right `N` when a generative model does the searching, and where does getting it wrong actually change decisions?

---

2. Background

Deflated Sharpe Ratio. Bailey and López de Prado (2014) define the benchmark Sharpe a best-of-`N` selection must beat as `SR*_0 = sqrt(V) * E[max_N]`, where `V` is the cross-trial variance of Sharpe estimates and `E[max_N]` is the expected maximum of `N` standard normals, approximated by `(1-gamma)*Z^{-1}(1-1/N) + gamma*Z^{-1}(1-1/(N e))` with `gamma` the Euler-Mascheroni constant. The DSR is the probabilistic Sharpe of the selected strategy against this inflated benchmark.

Probability of backtest overfitting. Bailey, Borwein, López de Prado and Zhu estimate, via combinatorially symmetric cross-validation (CSCV), the probability that the in-sample winner underperforms the median out-of-sample. It is model-free but still frames trials as distinct.

Effective number of tests. In statistical genetics, correlated tests across linked loci are handled by an *effective* number of independent tests `M_eff`, recovered from the eigenvalues of the test correlation matrix (Cheverud 2001; Li and Ji 2005; Galwey 2009). When tests are correlated, `M_eff < m`, and the multiple-testing correction is accordingly less severe. To my knowledge this device has not been carried into the finance or LLM-search setting.

---

3. Method

Given `m` candidate strategies with per-bar return series, form the `m x m` return-correlation matrix `C` and take its eigenvalues `lambda_i` (clipped at zero). I use Galwey's estimator as the default for its good behavior at the endpoints:

```
M_eff = (sum_i sqrt(lambda_i))^2 / sum_i lambda_i
```

which equals `m` at zero correlation and tends to 1 as correlation tends to 1. `M_eff` then replaces the raw count `m` wherever a trials count is required, in the DSR's expected-max term and in the practitioner log-bar. Li and Ji's estimator gives materially the same numbers throughout (Section 6). I test the estimator's own stability in Section 5.5 and find it near-unbiased and tight at realistic sample lengths, needing no shrinkage.

4. Where the correction bites

The magnitude of the correction depends entirely on how fast the gate's penalty grows in `N`.

- **Practitioner log-bar (first-order).** ÀYÒ's gate requires `1 + 0.15 * log2(N)`. Going from `N = m = 40` to `N = M_eff = 8` moves the required out-of-sample Sharpe from 1.80 to about 1.45, a decision-relevant 0.35.
- **Deflated Sharpe Ratio (second-order).** The DSR benchmark grows as `sqrt(2 ln N)`. From 40 to 8 that is roughly a 25 percent change in the benchmark, and the DSR-as-gate is in any case intrinsically conservative at realistic sample sizes (Section 5.2). The count barely moves its decisions.

The correction is therefore most useful for the log-style gates practitioners actually run, and largely cosmetic inside the DSR.

---

5. Synthetic experiments

All experiments use `T = 252` daily bars and `m = 40` proposals with a controllable equicorrelation `rho`, built from one common factor: `r_i = sqrt(rho) * f + sqrt(1−rho) * e_i`. Every strategy is pure noise unless an edge is explicitly injected.

5.1 The correction tracks proposal correlation

`M_eff` falls from about 38 at `rho = 0` to about 8 at `rho = 0.9`, while the raw count stays pinned at 40. A fixed count cannot represent the changing redundancy of LLM proposals; `M_eff` does.

5.2 Null calibration, and an honest negative result

Under the operational rule "deploy if DSR > 0.95" (nominal 5 percent false-deploy size), the actual false-deploy rate under pure noise:

rho	M_eff	N=1	N=m	N=M_eff
0.0	38.4	100.0%	0.0%	0.0%
0.3	31.5	99.8%	0.7%	0.9%
0.6	21.2	92.0%	2.5%	4.2%
0.9	8.2	41.1%	5.1%	7.6%

Mean absolute deviation from nominal: `N=1` 81.4 percent, `N=m` 3.1 percent, `N=M_eff` 3.0 percent. The lesson is twofold and I report both halves. First, correcting at all is essential: `N=1` deploys noise almost always. Second, and honestly, `M_eff` and `m` are nearly tied here, because the DSR test is intrinsically conservative at these sample sizes for reasons unrelated to the count. The selection winner fluctuates with standard deviation about 0.030 around the benchmark, but the DSR judges it against the single-strategy standard error of about 0.063, so it rarely fires for any count. I verified that the expected-max approximation is not the cause (asymptotic 0.1379 vs exact 0.1361). The DSR is simply the wrong place for the effective-count correction to show up.

5.3 Impact on the practitioner gate, with a decision flip

Where the penalty is first-order, the correction changes decisions. A genuine candidate with out-of-sample Sharpe 1.6 arrives:

rho	M_eff	bar(m=40)	bar(M_eff)	m verdict	M_eff verdict
0.0	38.4	1.80	1.79	REJECT	REJECT
0.3	31.5	1.80	1.75	REJECT	REJECT
0.6	21.2	1.80	1.66	REJECT	REJECT
0.85	10.6	1.80	1.51	REJECT	**DEPLOY**

At high proposal correlation, the regime LLM search actually produces, a real strategy the raw-count bar rejects is correctly deployed once effective trials are counted.

5.4 Why not just compare power at matched size

I also ran a matched-size power test: calibrate each count to reject noise at exactly 5 percent, then measure how often each deploys an injected edge. The three counts came out within a few points of each other. This is expected, not disappointing: once size is matched, the three DSRs are monotone re-thresholds of the same underlying winner statistic, so they trace nearly the same ROC curve. The count does not change the achievable power; it changes where your nominal operating point lands. That is precisely why the effect surfaces as calibration and as the position of a fixed log-bar, not as raw power.

5.5 Stability of the estimator

A trial count is only useful if it is stable at the sample sizes practitioners actually have. I Monte-Carlo the Galwey estimator on equicorrelated batches (`m = 14` , `rho = 0.6` , population `M_eff = 8.94`) across sample lengths and shrinkage intensities:

T	raw	shrink 0.1	shrink 0.2
126	8.8 ± 0.4	9.6 ± 0.3	10.3 ± 0.3
252	8.9 ± 0.3	9.6 ± 0.3	10.3 ± 0.2
504	8.9 ± 0.2	9.7 ± 0.2	10.4 ± 0.1

The raw estimator is near-unbiased and tight even at half a year of daily data (`8.8 ± 0.4` against a target of `8.94`), and the bias and variance both shrink with `T` as expected. Contrary to my initial expectation, shrinking the correlation matrix toward the identity is counterproductive: it biases `M_eff` upward toward `m`, because pulling correlations toward zero makes the candidates look more independent than they are. I therefore use the raw estimator with no shrinkage; the method has no tuning knob.

---

6. Empirical study on ÀYÒ

I asked the ÀYÒ research brain (the on-box `claude` CLI) for ten strategies across twelve deliberately diverse angles (momentum, mean reversion, breakout, quality-and-trend, reversal, volume confirmation, low-volatility drift, and others), and backtested each on one shared ten-name universe over 910 real daily bars from Alpaca. I then measured the return-correlation matrix and its effective rank. Code: `research/measure_ayo_correlation.py`.

Result. The ten proposals had an average pairwise return correlation of **+0.32** and an effective count of **M_eff = 7.3** (Galwey) / 7.0 (Li and Ji) against a raw count of `m = 10`. The redundancy is also visible directly in the rule vocabulary: nine of the ten proposals used the `sma_20_over_sma_50` trend filter and most also used `rsi_14`, despite the diverse prompt angles. The model draws repeatedly from a small set of indicators, exactly the behavior that makes the raw count an over-statement of independent trials.

Impact. At this batch size the effect on the log-bar is modest: `1 + 0.15*log2(10) = 1.50` versus `1 + 0.15*log2(7.3) = 1.43`, a 0.07 Sharpe reduction with no deployment flips at realistic out-of-sample Sharpe levels. Because the log-bar penalizes the redundancy *ratio* `m / M_eff` (here about 1.37), and that ratio is a property of the proposal generator rather than the batch size, the correction stays near 0.07 Sharpe even as `m` grows, unless the ratio itself grows.

Signal-space redundancy, and a negative result I expected to go the other way. I hypothesized that return-stream correlation understates redundancy, since strategies encoding the same idea might trade at different times, and that measuring holdings overlap (are two strategies *in a position on the same days*) would reveal a lower, truer `M_eff`. On a fresh 12-strategy batch I measured both on the same proposals. The hypothesis was refuted: signal-space average correlation was **+0.07** with `M_eff = 10.8`, *higher* (more independent) than return-space (+0.23, `M_eff = 9.1`), against `m = 12`. Shared vocabulary is not shared behavior: the proposals reuse the same indicators but with opposite timing (a

momentum rule holds in uptrends, a mean-reversion rule after drops, a fade-overbought rule at RSI peaks), so they are rarely in the market together.

A controlled contrast: inducing redundancy. To test whether `M_eff` actually catches redundancy when it is present, I repeated the measurement but drew fourteen proposals from a *single narrow angle* (oversold RSI mean reversion) rather than twelve diverse ones. The model converged hard: nearly every proposal was a variant named "Oversold RSI Snapback." The numbers followed. Average return correlation rose to **+0.72** and signal-space to **+0.78** (the two spaces now agree, confirming genuine redundancy in both idea and holdings), giving `M_eff = 3.0` against `m = 14`, a **4.6x** over-count. The trials-adjusted bar drops from 1.57 to 1.24, and a candidate with out-of-sample Sharpe 1.3 that raw counting *rejects* (as one of 14 trials) is correctly *deployed* once counted as one of ~3 effective trials.

generator	avg corr (return)	M_eff	m/M_eff	bar
diverse angles (ÀYÒ's protocol)	+0.23	9.1	1.3x	1.48
single narrow angle	+0.72	3.0	4.6x	1.24

Conclusion. The magnitude of the effective-trials correction is a property of the *generation protocol*, not a universal constant, and I demonstrate both poles on one real system. A diverse protocol produces near-independent proposals (`M_eff` approximately `m`) and needs no correction; a redundant protocol over-counts by 4 to 5 times, inflates the required Sharpe by roughly 0.33, and mis-rejects genuine strategies. `M_eff` is the instrument that identifies which regime a given generator is in and applies a correction only when one is warranted. For ÀYÒ specifically it certifies that the diverse-angle proposal design keeps the existing raw-count bar close to correct, which is itself a useful negative result: prompt diversity is not cosmetic, it is what keeps the selection statistics honest.

---

7. Limitations

- Return correlation is linear; strategy dependence may not be. A rank-correlation variant is worth checking, though the return- and signal-space measurements agreeing in the redundant regime is reassuring.
- The effect is concentrated where the proposal generator is redundant. A diverse generator gives `M_eff` approximately `m` and there is little to correct, as I found for ÀYÒ's own protocol.
- Proposals arriving adaptively (bandit-style) break the fixed-batch assumption; I scope to the batch case.
- The residual conservatism of the DSR-as-gate is a separate, count-independent phenomenon and is not addressed here.
- I evaluate on one strategy family space (typed technical rules over a fixed indicator vocabulary). Whether the protocol-dependence generalizes to richer proposal spaces is open.

8. Conclusion

When a language model is the search process, the trial count that selection-bias corrections depend on is not the number of proposals but the effective number of independent ones. Borrowing the effective-number-of-tests estimator from genetics gives a principled `M_eff` that is near-unbiased and stable at realistic sample sizes. Its impact is first-order in the trials-adjusted bars practitioners actually run and second-order in the Deflated Sharpe Ratio. Most importantly, on a real LLM-driven quant pipeline the size of the correction turns out to be a property of the generation protocol, not a universal constant: a diverse prompting protocol produces near-independent proposals that need no correction, while a narrow one collapses fourteen proposals to three effective trials and mis-rejects genuine strategies. `M_eff` is the instrument that tells you which regime you are in, and it certifies, as a useful negative result, that prompt diversity is what keeps a generative strategy search statistically honest.

References

- Bailey, D. H., & López de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality. *Journal of Portfolio Management*.
- Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The Probability of Backtest Overfitting. *Journal of Computational Finance*.
- Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*.
- Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*.
- Galwey, N. W. (2009). A new measure of the effective number of tests. *Genetic Epidemiology*.