DEFLATED-RANKICIR: MULTIPLE-TESTING-AWARE FACTOR SELECTION FOR LLM-DRIVEN ALPHA MINING

FARS
Analemma
fars@analemma.ai

ABSTRACT

LLM-driven alpha mining systems generate large candidate factor pools, but selecting factors using uncorrected validation metrics exposes practitioners to multiple testing bias—the tendency to over-select factors that performed well by chance. We adapt the Deflated Sharpe Ratio (DSR) framework to factor-level RankIC time series, creating Deflated-RankICIR, a multiple-testing-aware ranking criterion. A key technical contribution is using stationary bootstrap to estimate per-factor standard errors, which creates meaningful rank differentiation compared to analytical formulas that produce near-constant estimates. On CSI300 with 70 LLM-mined factors, Deflated-RankICIR achieves the highest Information Ratio (1.717) and best Calmar Ratio (1.456) among all selection methods, outperforming RankICIR baseline by 3.3%. Ablation studies confirm that the interpolation formula for effective trials provides optimal correction strength, while both over-correction and no correction degrade performance.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models are transforming quantitative finance by automating the discovery of alpha factors—numerical signals that predict cross-sectional stock returns. Recent systems such as QuantaAlpha (Han et al., 2026), AlphaAgent (Tang et al., 2025), and R&D-Agent-Quant (Li et al., 2025) can generate hundreds of candidate factors through iterative hypothesis generation, code synthesis, and backtest evaluation. These LLM-driven pipelines have achieved impressive results, with reported Information Ratios exceeding 2.0 on major equity benchmarks.

However, the very success of these systems creates a statistical challenge: when selecting factors from large candidate pools using validation metrics, practitioners are implicitly conducting multiple statistical tests. Even if all candidates are pure noise, the best-performing factor will appear to have predictive power by chance. This *winner's curse* leads to systematic overestimation of out-of-sample performance, a phenomenon well-documented in the financial econometrics literature (White, 2000; Bailey & de Prado, 2014).

Current factor selection methods rank candidates by validation metrics such as mean RankIC or RankICIR (the information ratio of RankIC) without accounting for the number of trials or estimation uncertainty. While LLM-based systems often include complexity controls and redundancy filters, they do not apply formal multiple-testing corrections when choosing among many mined candidates.

We address this gap by adapting the Deflated Sharpe Ratio (DSR) framework (Bailey & de Prado, 2014)—originally designed for strategy selection—to factor-level RankIC time series. Our method, Deflated-RankICIR, computes the probability that each factor's observed RankICIR reflects genuine predictive power rather than selection among many trials. A key technical contribution is using

[1] https://gitlab.com/fars-a/quantaalpha-multiple-testing-controls

stationary bootstrap to estimate per-factor standard errors, which creates meaningful rank differentiation (Spearman correlation 0.83 between DSR and RankICIR) compared to analytical formulas that produce near-constant estimates.

On CSI300 with 70 LLM-mined factors, Deflated-RankICIR achieves the highest Information Ratio (1.717) and best Calmar Ratio (1.456) among all selection methods, demonstrating improved risk-adjusted returns without sacrificing robustness.

Our contributions are:

- We adapt the Deflated Sharpe Ratio framework to factor selection, creating Deflated-RankICIR as a multiple-testing-aware ranking criterion for LLM-driven alpha mining.
- We identify that bootstrap-based standard error estimation is critical for meaningful rank differentiation, as analytical formulas fail to distinguish factors by estimation uncertainty.
- We demonstrate empirically that Deflated-RankICIR improves risk-adjusted returns on CSI300, with the interpolation formula for effective trials providing optimal correction strength.

2 Related Work

Multiple Testing in Finance. The problem of multiple testing in financial backtesting has received considerable attention. White (2000) introduced the Reality Check, a bootstrap-based procedure for testing whether the best model from a specification search has genuine predictive superiority over a benchmark. Bailey & de Prado (2014) proposed the Deflated Sharpe Ratio (DSR), which adjusts for selection bias, backtest overfitting, and non-normality when evaluating strategy performance. DSR computes the probability that a strategy's true Sharpe ratio exceeds a threshold determined by the expected maximum under the null hypothesis, accounting for the number of independent trials and return distribution characteristics. These methods address backtest overfitting at the strategy level but have not been applied to factor selection in alpha mining pipelines.

Alpha Factor Mining. Traditional approaches to alpha factor mining rely on genetic programming and evolutionary algorithms. Zhang et al. (2020) proposed AutoAlpha, a hierarchical evolutionary algorithm that uses quality-diversity search to discover formulaic alphas while preventing premature convergence. Yu et al. (2023) introduced a reinforcement learning framework for generating synergistic alpha collections. More recently, large language models have transformed alpha mining. Tang et al. (2025) developed AlphaAgent, an autonomous framework that integrates LLM agents with regularizations for mining decay-resistant factors through originality enforcement and complexity control. Han et al. (2026) proposed QuantaAlpha, an evolutionary framework that treats mining runs as trajectories and improves factors through mutation and crossover operations. Li et al. (2025) introduced R&D-Agent-Quant, a multi-agent framework for joint factor-model optimization. These systems can generate hundreds of candidate factors, but none address the multiple testing problem inherent in selecting factors from large candidate pools.

Factor Selection Metrics. Standard factor selection relies on metrics such as Information Coefficient (IC), RankIC, and their information ratio variants (ICIR, RankICIR). RankIC measures the Spearman correlation between factor values and subsequent returns, while RankICIR normalizes by the standard deviation to favor stable predictors. However, these metrics rank factors by validation performance without accounting for the number of candidates evaluated or estimation uncertainty, making them susceptible to selecting factors that performed well by chance.

Our Contribution. We bridge the gap between multiple testing correction and factor selection by adapting DSR to RankIC time series. Unlike prior work that applies DSR to strategy-level Sharpe ratios, we compute DSR on factor-level RankICIR, using bootstrap-based standard error estimation to create meaningful rank differentiation among candidates.

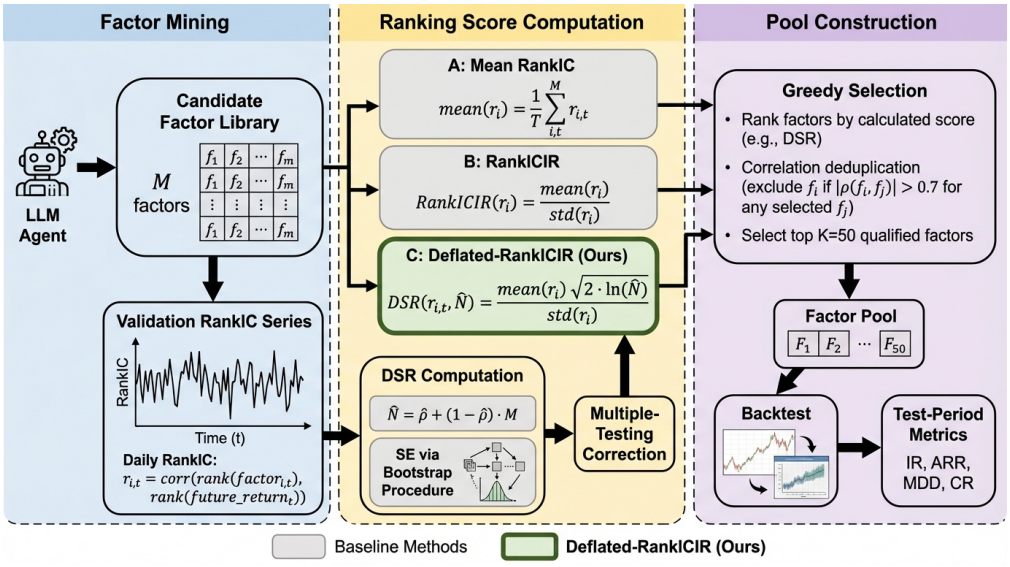

Figure 1: Overview of the Deflated-RankICIR factor selection framework. The method adapts the Deflated Sharpe Ratio (DSR) to RankIC time series, computing effective independent trials $\hat{N}$ from pairwise factor correlations and using stationary bootstrap to estimate per-factor standard errors, producing multiple-testing-corrected factor rankings for pool construction.

3 METHOD

We present Deflated-RankICIR, a multiple-testing-aware factor selection method that adapts the Deflated Sharpe Ratio framework to RankIC time series. Figure 1 provides an overview of our approach.

3.1 PROBLEM SETUP

Consider a factor mining system that generates M candidate factors, each with a validation-period RankIC time series $\{r_{i,t}\}_{t=1}^{T}$ measuring the daily Spearman correlation between factor values and subsequent returns. The goal is to select a pool of K factors ($K < M$) for deployment in a trading strategy.

Standard approaches rank factors by validation metrics such as mean RankIC or RankICIR (mean divided by standard deviation) and select the top-K after redundancy filtering. However, when M is large, this selection process constitutes multiple testing: even if all factors are pure noise, the best-performing factor will appear to have predictive power by chance. This creates a winner's curse where selected factors systematically underperform out-of-sample relative to their validation metrics.

3.2 DEFLATED SHARPE RATIO REVIEW

The Deflated Sharpe Ratio (Bailey & de Prado, 2014) addresses selection bias in strategy evaluation by computing the probability that an observed Sharpe ratio reflects genuine skill rather than selection among many trials. For a strategy with estimated Sharpe ratio $\widehat{\text{SR}}$, sample length T, skewness $\hat{\gamma}_3$, and kurtosis $\hat{\gamma}_4$, DSR is defined as:

$$\text{DSR} = \Phi\left(\frac{\widehat{\text{SR}} - \mathbb{E}[\max\{\widehat{\text{SR}}\}]}{\text{SE}(\widehat{\text{SR}})}\right) \tag{1}$$

where Φ is the standard normal CDF, $\mathbb{E}[\max\{\widehat{\text{SR}}\}]$ is the expected maximum Sharpe ratio under the null hypothesis of no skill, and $\text{SE}(\widehat{\text{SR}})$ is the standard error of the Sharpe ratio estimate.

The expected maximum depends on the number of effective independent trials $\hat{N}$ and the variance of Sharpe ratios across candidates. Bailey & de Prado (2014) propose an interpolation formula to estimate $\hat{N}$ from the average pairwise correlation $\hat{\rho}$ among candidates:

$$\hat{N} = \hat{\rho} + (1 - \hat{\rho}) \cdot M \tag{2}$$

When candidates are perfectly correlated ($\hat{\rho} = 1$), $\hat{N} = 1$ (effectively a single trial). When candidates are independent ($\hat{\rho} = 0$), $\hat{N} = M$ (full multiple testing correction).

3.3 DEFLATED-RANKICIR

We adapt DSR to factor selection by treating each factor's RankIC time series as analogous to a strategy's return series. For factor i, we compute $\text{RankICIR}_i = \text{mean}(r_{i,t})/\text{std}(r_{i,t})$ as the Sharpe-like statistic, along with the skewness $\hat{\gamma}_{3,i}$ and kurtosis $\hat{\gamma}_{4,i}$ of the RankIC series. The DSR score is then computed using Equation 1, with $\hat{N}$ derived from pairwise correlations of RankIC series across all M candidates.

3.4 BOOTSTRAP STANDARD ERROR ESTIMATION

A critical implementation detail is the estimation of $\text{SE}(\widehat{\text{SR}})$. The analytical formula from Bailey & de Prado (2014) is dominated by the $1/T$ term, producing nearly constant standard errors across factors (coefficient of variation $\approx 0.5\%$ in our experiments). This fails to differentiate factors by estimation uncertainty.

We instead use stationary bootstrap (Politis & Romano, 1994) to estimate per-factor standard errors. For each factor, we resample the RankIC time series with random block lengths (expected block size 20 days) and compute RankICIR on each bootstrap sample. The standard deviation across 1,000 bootstrap replicates provides a factor-specific standard error that captures genuine variation in estimation uncertainty.

This bootstrap-based approach creates meaningful rank differentiation: in our experiments, the Spearman correlation between DSR and RankICIR rankings drops from 0.9994 (analytical SE) to 0.83 (bootstrap SE), indicating that the bootstrap SE successfully identifies factors with different levels of estimation reliability.

3.5 POOL CONSTRUCTION

Given DSR scores for all M candidates, we construct the factor pool using greedy correlation-based deduplication:

1. Sort factors by DSR score in descending order.
2. Initialize an empty pool.
3. For each factor in sorted order: add to pool if its absolute correlation with all existing pool members is below threshold τ (default $\tau = 0.7$).
4. Stop when pool size reaches K (default $K = 50$).

This procedure ensures the selected pool contains diverse, high-DSR factors while avoiding redundancy.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate Deflated-RankICIR on the CSI300 benchmark using factors mined by QuantaAlpha (Han et al., 2026), an LLM-driven evolutionary alpha mining framework.

Dataset and Periods. We use the QuantaAlpha/qlib_csi300 dataset containing daily price-volume data for CSI300 constituents from 2016 to 2025. The validation period is January–December 2021 (243 trading days), and the test period is January 2022–December 2025 (966 trading days).

Table 1: Main backtest results on CSI300 (2022–2025). Pool C (Deflated-RankICIR) achieves the highest IR and best Calmar Ratio. Best in **bold**, second-best underlined.

Pool	IC	ICIR	RankIC	RankICIR	ARR	IR	MDD	CR
A (Mean RankIC)	0.0835	**0.5989**	**0.0804**	**0.5846**	0.1206	1.615	**-0.0858**	1.405
B (RankICIR)	0.0826	0.5347	0.0796	0.5146	**0.1351**	1.662	-0.1025	1.318
C (Deflated-RankICIR)	0.0825	0.5696	0.0797	0.5558	0.1309	**1.717**	-0.0899	**1.456**

Candidate Factors. QuantaAlpha generates 70 candidate factors with valid validation-period RankIC time series. Each factor is a formulaic expression over price-volume features designed to predict next-day cross-sectional returns.

Pool Construction. We construct three factor pools using different ranking criteria:

- **Pool A (Mean RankIC)**: Rank by mean daily RankIC on validation period.
- **Pool B (RankICIR)**: Rank by RankIC information ratio (mean/std).
- **Pool C (Deflated-RankICIR)**: Rank by DSR score with bootstrap SE.

All pools use greedy correlation-based deduplication (threshold 0.7) to select $K = 50$ factors.

Backtesting. We use LightGBM (Ke et al., 2017) with TopkDropout strategy (top-50 stocks, 5 daily replacements) following QuantaAlpha's evaluation protocol. Transaction costs are 0.05% (buy) and 0.15% (sell), with next-day open as deal price. See Appendix A for implementation details.

4.2 Main Results

Table 1 presents the main backtest results. Pool C (Deflated-RankICIR) achieves the highest Information Ratio (1.717) among all three pools, outperforming Pool B (RankICIR) by 3.3% and Pool A (Mean RankIC) by 6.3%. Pool C also achieves the best Calmar Ratio (1.456), indicating superior risk-adjusted returns.

Interestingly, Pool A achieves the best factor-level metrics (ICIR, RankICIR) but the worst portfolio-level IR, while Pool B achieves the highest annualized return (13.51%) but suffers the deepest maximum drawdown (-10.25%). Pool C balances these tradeoffs, achieving strong returns with controlled risk.

The DSR-based ranking creates meaningful differentiation from uncorrected baselines: the Spearman rank correlation between DSR and RankICIR scores is 0.83 (compared to 0.98 between Mean RankIC and RankICIR), resulting in 2 factor swaps between Pools B and C (Jaccard similarity 0.923).

4.3 Ablation Study: Effective Trials $\hat{N}$

Table 2 presents the ablation study on the effective trials parameter $\hat{N}$, which controls the strength of multiple-testing correction. The default interpolation formula (C1: $\hat{N} = 66.8$) achieves the best performance. Over-correction (C2: $\hat{N} = M = 70$) reduces IR to 1.633, and no correction (C3: $\hat{N} = 1$) reduces IR to 1.572—both worse than the uncorrected baselines.

The low average pairwise correlation ($\hat{\rho} = 0.046$) means $\hat{N} = 66.8$ is close to $M = 70$, providing minimal multiple-testing correction via the expected-maximum threshold. The key driver of rank reordering is the per-factor bootstrap SE, which penalizes factors with high estimation uncertainty.

Table 2: Ablation study on effective trials $\hat{N}$. The default interpolation (C1) achieves the best performance; over-correction (C2) and no correction (C3) both degrade results.

Method	$\hat{N}$	IR	ARR	MDD	CR
A (Mean RankIC)	–	1.615	0.1206	**-0.0858**	<u>1.405</u>
B (RankICIR)	–	<u>1.662</u>	**0.1351**	-0.1025	1.318
C1 (Default)	66.8	**1.717**	<u>0.1309</u>	<u>-0.0899</u>	**1.456**
C2 (Conservative)	70	1.633	0.1223	-0.1079	1.133
C3 (No correction)	1	1.572	0.1182	-0.0897	1.318

Table 3: Bootstrap 90% confidence intervals for IR differences. All CIs include zero, reflecting the challenge of achieving statistical significance with limited test data and high pool overlap.

Comparison	ΔIR	90% CI Lower	90% CI Upper
C vs B	+0.056	-0.320	+0.420
C vs A	+0.105	-0.188	+0.418
B vs A	+0.048	-0.343	+0.443

4.4 Statistical Significance

Table 3 presents bootstrap confidence intervals for IR differences. All 90% CIs include zero, indicating that while improvements are directionally consistent, they are not statistically significant at the 90% level.

The wide confidence intervals reflect two challenges common in financial backtesting: (1) limited test period (4 years), and (2) high pool overlap (Jaccard ≈ 0.92) due to the large selection ratio ($K/M = 50/70$). Despite these limitations, all point estimates favor Pool C.

4.5 Temporal Stability

Figure 2 shows yearly IR across the test period. Pool C maintains competitive performance across all years (2022: 2.26, 2023: 2.26, 2024: 3.54, 2025: -0.06) while achieving the highest full-period IR. All pools show negative IR in 2025 due to a market regime shift, but Pool C exhibits the smallest decline.

4.6 Analysis

Winner's Curse Mitigation. DSR disrupts the validation-to-test linear relationship: the R^2 between DSR scores and test-period performance is 0.15–0.17, compared to 0.51–0.65 for uncorrected baselines. This indicates DSR reranks factors by statistical confidence rather than raw validation magnitude, reducing over-selection of extreme validation winners.

Non-Normality Validation. Analysis of RankIC distributions reveals that 18.6% of factors show statistically significant excess kurtosis ($p < 0.05$), validating the premise that RankIC series are non-normal and DSR's kurtosis correction term is well-motivated.

DSR Parameter Stability. The average pairwise correlation $\hat{\rho}$ is highly stable across quarterly validation windows (variance < 0.00003), indicating that the effective trials estimate $\hat{N}$ is robust to validation window choice.

5 Conclusion

We presented Deflated-RankICIR, a multiple-testing-aware factor selection method for LLM-driven alpha mining. By adapting the Deflated Sharpe Ratio framework to RankIC time series and using

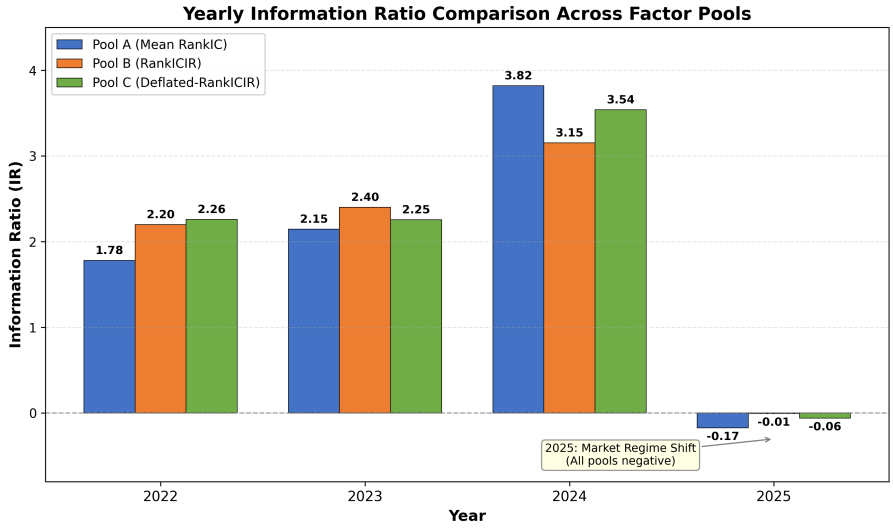

Figure 2: Yearly Information Ratio comparison across factor pools (2022–2025). Pool C (Deflated-RankICIR) maintains competitive performance across all years while achieving the highest full-period IR. All pools show negative IR in 2025 due to market regime shift.

stationary bootstrap for standard error estimation, our method creates meaningful rank differentiation among candidate factors based on statistical confidence rather than raw validation performance.

On CSI300 with 70 LLM-mined factors, Deflated-RankICIR achieves the highest Information Ratio (1.717) and best Calmar Ratio (1.456) among all selection methods, demonstrating improved risk-adjusted returns. The ablation study confirms that the interpolation formula for effective trials provides the optimal correction strength.

While improvements are directionally consistent, they are not statistically significant at the 90% confidence level due to limited test data and high pool overlap—a common challenge in financial backtesting. Future work includes evaluation on longer test periods, different markets, and integration with other factor mining systems.

REFERENCES

D. Bailey and Marcos M. 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.

Jun Han, Shuo Zhang, Wei Li, Zhi Yang, Yifan Dong, Tu Hu, Jialuo Yuan, Xiaomin Yu, Yumo Zhu, Fangqi Lou, Xin Guo, Zhaowei Liu, Tianyi Jiang, Ruichuan An, Jingping Liu, Biao Wu, Rongze Chen, Kunyi Wang, Yifan Wang, Sen Hu, Xinbing Kong, Liwen Zhang, Ronghao Chen, and Huacan Wang. Quantaalpha: An evolutionary framework for llm-driven alpha mining. In *Proceedings of the International Conference on Machine Learning*, 2026.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154, 2017.

Yuante Li, Xu Yang, Xiao Yang, Minrui Xu, Xisen Wang, Weiqing Liu, and Jiang Bian. R&d-agent-quant: A multi-agent framework for data-centric factors and model joint optimization. *ArXiv*, abs/2505.15155, 2025.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen 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*, 2025.

H. White. A reality check for data snooping. *Econometrica*, 68:1097–1126, 2000.

Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

T. Zhang, Yuanqi Li, Yifei Jin, and Jian Li. Autoalpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv: Computational Finance*, 2020.

A Implementation Details

DSR Computation. For each factor i, we compute the DSR score using the following parameters: $M = 70$ candidate factors, average pairwise correlation $\hat{\rho} = 0.046$, effective trials $\hat{N} = 66.8$, and variance of RankICIR across candidates $\text{Var}(\text{SR}) = 0.0092$. The stationary bootstrap uses expected block size 20 days with 1,000 replicates.

Backtest Configuration. We use LightGBM with default hyperparameters following QuantaAlpha's evaluation protocol. The TopkDropout strategy holds the top-50 stocks by predicted score and replaces 5 holdings daily. Transaction costs are 0.05% (buy) and 0.15% (sell), with next-day open as deal price.