The Backtest That Lied:

A Forensic Investigation of Leakage, Implementation Failure, and Liquidity Dependence in A-Share Machine Learning

Technical Research Brief

Student Researcher

July 2026

1 Abstract

A near-perfect cross-sectional prediction signal—monthly rank IC 0.940 collapsed to 0.100 after four detected data leakages were corrected. The surviving signal, mean OOS IC 0.108 with Newey–West $t = 10.5$, represented genuine low-liquidity stock-selection ability. With executable portfolios and transaction costs, the strategy produced negative absolute returns in the liquid U50 universe (net annual -0.9%, max drawdown 61.3%). The OHLCV-only hypothesis was formally terminated. The primary contribution is methodological.

2 Introduction

> *A model can predict returns and still fail as a strategy.*

Initial result: IC 0.940 across 96 consecutive OOS months. After four leakages, apparent predictive power collapsed by 89%. Retained signal: mean OOS IC 0.108, NW $t = 10.5$.

(1) Predictability: IC 0.108 survived null infrastructure.

(2) Concentration: 48.1% of holdings in Q1. Within-Q1: +2.7% monthly excess.

(3) Liquid universes: U50 net -0.9%, DD 61.3%.

(4) Longer horizons: H3 -21.7%, H6 -15.2%.

The OHLCV-only strategy was terminated. Offered as a research methodology case study.

3 Research Question

The primary research question:

> *Can daily OHLCV data produce a statistically valid, economically executable, and scalable A-share trading strategy?*

Three hierarchical tests:

T1: Is the predictive signal real? — Cross-sectional rank IC out-of-sample, after all leak corrections.

T2: Can the signal survive execution and costs? — Next-open execution, realistic costs, 90/80 buffer, absolute returns.

T3: Can the signal survive liquidity and capacity? — U50/U30 vs full universe, ADV20 quintile decomposition.

Each test evaluated against pre-registered gates with no post-hoc changes.

4 Data and Methodology

4.1 Data

RQAlpha A-share daily market-data bundle, 5511.0 historically listed stocks. Point-in-time universe without survivorship filters. Daily OHLCV aggregated to monthly. True PIT fundamental data (market cap, book value, earnings, industry) was unavailable.

4.2 Features

20 ranked technical indicators: momentum (5/10/20/60/120-day), volatility (5/20/60-day), volume (5/20-day turnover), MA deviation (20/60-day), channel position, RSI(14), turnover, reversal (1/5-day), cross-sectional rank, skewness, breadth. All rank-normalized to $[-1, 1]$ per month.

4.3 Model

LightGBM: 200 trees, lr=0.02, depth=3, leaves=8, subsample=0.7, colsample=0.7, L1=1, L2=10. 20-seed ensemble. No hyperparameter tuning.

4.4 Validation

Three walk-forward folds: Fold 1 (2010-16→2017-19), Fold 2 (2013-19→2020-21), Fold 3 (2015-21→2022-24). Inference: Spearman IC, Newey–West (lags 3,6), block bootstrap (1000 resamples, block 6 months). Null infrastructure: iid Gaussian null, constant-label test, permutation audit.

4.5 Portfolio Construction

Monthly rebalance, top 10% by rank, enter ≥90th, retain ≥80th. Next-open execution. Equal weight. Cost scenarios: Gross (0bp), Base (13bp buy/18bp sell), Stress (22.5bp/27.5bp).

4.6 Evidence Architecture

All numerical claims generated from `data/evidence.json` (39 metrics, 3 collapses). `scripts/validate_evidenc` + `tests/test_evidence.py` enforce integrity.

5 The Three Collapses

5.1 Collapse I: Statistical Illusion

IC 0.940 → 0.100 after fixing contemporaneous target leakage. 89% of apparent predictive power was mechanical illusion (absolute decline: 0.84 IC points).

5.2 Collapse II: Economic Implementation

IC 0.108 failed to produce positive U50 absolute returns: net -0.9%, DD 61.3%. Benchmark: -5.9%.

5.3 Collapse III: Scalability

U100 net +6.8% but 48.1% of holdings in Q1 (median ADV 19.0M¥). Within-Q1: +2.7% monthly excess. The alpha existed where capacity was severely constrained.

6 Portfolio Implementation

Table 1: Portfolio performance (31bp round-trip).

Portfolio	Net Annual	Sharpe	Max DD	Excess
U100 RAW (base31)	+6.8%	+0.31	46.3%	+6.9%
U50 RESID (base31)	-0.9%	-0.04	61.3%	+4.9%
U50 EW benchmark	-5.9%	-0.26	—	—

U50 RESID generates +4.9% annual excess over the benchmark (-5.9%).

7 Liquidity Attribution

Table 2: Liquidity quintile composition.

Quintile	Median ADV (M¥)	% Holdings	Net Annual
Q1 (lowest)	19.0	48.1%	+10.4%
Q2	41.0	25.8%	+3.9%
Q3	75.0	14.0%	-0.6%
Q4	145.0	7.8%	-3.5%
Q5 (highest)	386.0	4.2%	-2.6%

Within-Q1 selection: +2.7% per month. The signal was concentrated in the lowest-liquidity stocks and was severely capacity-constrained. Maximum feasible capacity was not fully quantified.

8 Residual and Horizon Tests

8.1 Cross-Fitted Residual Target (RESID)

Own-stock contamination: IC 0.760 → 0.120 after cross-fitting. Residualization improved U50 marginally (net -0.9%, DD 61.3%) but could not reach profitability.

8.2 Multi-Horizon Targets (H3 and H6)

H3 U50: -21.7% annual. H6 U50: -15.2% annual. Multi-horizon targeting does not rescue OHLCV-only scalability.

9 The Signal Was Real. The Strategy Was Not.

Frozen baseline: IC 0.108, NW $t = 10.5$. Pre-registered gates prevented post-hoc optimization. Termination was the only honest decision.

Table 3: Decision gates.

Criterion	**Result**	**Gate**
Predictive ranking	Passed	IC > 0, NW $t > 2$
Time-aware significance	Passed	Bootstrap CI excludes 0
Within-Q1 selection	Passed	Excess > 0
Positive U50 return	Failed	Net < 0
Acceptable drawdown	Failed	DD $> 25\%$
Liquid-universe scalability	Failed	U50/U30 negative
Decision	**TERMINATED**	OHLCV branch closed

10 Limitations

(1) **Data scope:** Only daily OHLCV. No PIT fundamental data (market cap, book value, earnings, industry).

(2) **No market cap:** ADV20 findings are liquidity effects, not size effects. PIT market cap unavailable.

(3) **Non-redistributable data:** RQAlpha bundle is proprietary. `evidence.json` is the authoritative publication layer.

(4) **Sample consumption:** 2022–2024 used as validation; genuine forward test requires post-July 2026 data.

(5) **Cost model:** Flat per-trade costs. True Q1 costs likely higher.

(6) **Model family:** Only LightGBM tested. Other models may extract different information.

(7) **Factor breadth:** 20 technical factors only. No factor discovery or alternative data.

(8) **No capacity model:** Formal market-impact model not constructed. Capacity estimates approximate.

(9) **Survivorship:** PIT universe includes delisted stocks; delisting returns not explicitly modeled.

11 Conclusion

OHLCV features contain genuine ranking information: IC 0.108, NW $t = 10.5$, within-Q1 selection +2.7% monthly excess. The signal did not survive executable portfolios: U50 net -0.9%, DD 61.3%. Multi-horizon: H3 -21.7%, H6 -15.2%. The contribution is a demonstration that clean negative results are more valuable than profitable-looking lies.

A Evidence Map

Key metrics (of 39 total):

- `target_leak_before_ic`: IC 0.940 → `EXP_TARGET_SHIFT_01`
- `frozen_baseline_mean_ic`: IC 0.108 → `EXP_BASELINE_V43`
- `u100_raw_net_annual`: +6.8% → `EXP_DP1_PORTFOLIO`

- `u50_resid_net_annual`: -0.9% → `EXP_STEP5_ABLATION`
- `q1_within_excess`: +2.7% → `EXP_DP1A_QUINTILE`
- `h3_u50_net_annual`: -21.7% → `EXP_MH1R_HORIZON`

B Experiment Ledger

1. **Target Timing**: IC 0.940 → 0.100 (89% collapse). Invalid → Validated.
2. **Cross-Fit Audit**: IC 0.760 → 0.120. Invalid → Validated.
3. **Frozen Baseline**: Mean OOS IC 0.108, NW $t = 10.5$. Validated.
4. **Null Infrastructure**: Permutation correlation ≈ 0. Validated.
5. **Portfolio Diagnostics**: U100 +6.8%, U50 -0.9%. Q1-driven.
6. **Residual Ablation**: Cannot reach profitability. Rejected.
7. **Multi-Horizon**: H3 -21.7%, H6 -15.2% U50 net. Rejected.
8. **Termination**: OHLCV-only core strategy closed. Recorded.

C Reproducibility

C.1 Fully Reproducible

Evidence validation, website rendering, and all report metrics are reproducible from the repository with `python report/build.py`.

C.2 Not Fully Reproducible

Model training requires the RQAlpha market-data bundle (∼3.3GB, proprietary). Training scripts are included but data-dependent. `data/evidence.json` serves as the authoritative publication layer.

C.3 Repository

`https://github.com/DresdenGman/the-backtest-that-lied` — See repository tags for version information.

References

[1] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5), 2223–2273.

[2] Leippold, M., Wang, Q., & Zhou, W. (2022). Machine Learning in the Chinese Stock Market. *Journal of Financial Economics*, 145(2), 64–82.

[3] Liu, J., Stambaugh, R. F., & Yuan, Y. (2019). Size and Value in China. *Journal of Financial Economics*, 134(1), 48–69.

[4] Gârleanu, N., & Pedersen, L. H. (2013). Dynamic Trading with Predictable Returns and Transaction Costs. *Journal of Finance*, 68(6), 2309–2340.

[5] Lo, A. W., & MacKinlay, A. C. (1990). Data-Snooping Biases in Tests of Financial Asset Pricing Models. *Review of Financial Studies*, 3(3), 431–467.

[6] Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-Mathematics and Financial Charlatanism. *Notices of the AMS*, 61(5), 458–471.

[7] Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708.

[8] Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. *Monographs on Statistics and Applied Probability*, 57.