

Placebo-Audited LightGBM with Size-Block Covariance

Xinhe (Ashley) Wen, Yiting Huang, Bingyan (Skye) Wang, Emily Wu

Abstract. Can explicit size-group estimation add value beyond a pooled machine-learning model, and in which layer? We test size-group structure in the forecasts and the risk model of a LightGBM portfolio with an adaptive exposure overlay. In our experiments it helps in the risk model, not in the forecasts: a Ledoit–Wolf target that pools correlations by pairs of size groups adds 0.15 to the design-period Sharpe ratio, and inside the final model it beats all 30 random-group controls. Group-specific forecast corrections and a committee-disagreement penalty lowered the design-period Sharpe ratio; both remain in the submitted model. On the test period (408 monthly returns, January 1990 to December 2023), the final model has a Sharpe ratio of 3.58,¹ with a maximum drawdown of 8.4% and a gross Sharpe ratio of 1.3 in 2020–2023.

1 Research question and contribution

Machine-learning skill in stock returns differs by firm size: on our design data, our reference model (M3) earns a Sharpe ratio of about 4.5 in small caps and about 2.0 in mega caps. A pooled model already sees firm size through its inputs, so we ask whether separate estimation by size group adds value. We use partial pooling, the core idea of multilevel models: each group departs from the common estimate only as far as its data support. We test it in the forecasts and the risk model and find that it helps only in the risk model. Our scan of the code and docstrings of 47 public leaderboard entries did not identify size-block risk, staggered refit vintages or placebo tests of a model’s own components.

2 Data and research protocol

We use only the competition data: 1.37 million stock-months (1951–2023) with 402 characteristics, and 103 million daily excess returns. Each month’s weights use only data available at its end: cutting the data at December 2005, with unrealized targets removed, leaves every earlier weight unchanged. Three model vintages refit every three years, offset by 12 months, so one updates each year (Figure 1). Refitting adds new data; it never changes the design.

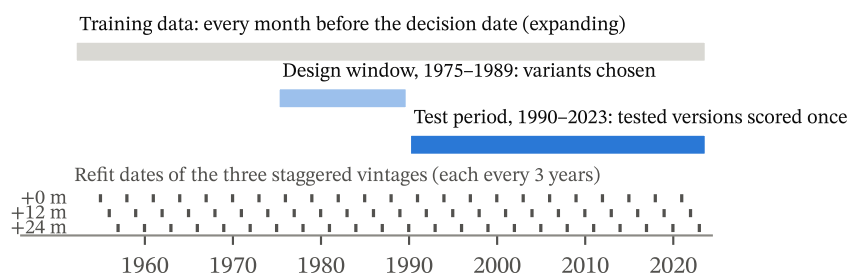


Figure 1. Variants were chosen on 1975–1989; tested versions were scored once on 1990–2023.

Source: JKP CTF data; model configuration of the submitted code.

Research protocol. We developed the model in versions, each pre-registered from M2 on: before a version ran, we recorded and hashed its candidate variants, gates and selection rule. Variants were compared with their parent on the design window, January 1975 to November 1989, and had to pass five gates: significant improvement (Newey–West $t \geq 2$, 6 lags), improvement in both halves, no fading, honest risk forecasts, and breadth across size groups. From M7 on, we ranked creativity ahead of Sharpe ratio, so the first gate required

¹All returns in this report are hypothetical backtests, gross of trading costs, scaled to 10% annual volatility over the full period as on the leaderboard.

only that every Sharpe ratio, overall, by half and by size group, stay above $3.0/3.58 = 84\%$ of the parent's, where 3.58 is M3's test Sharpe ratio. The selected variant was then scored once on the test period, 1990–2023. Test scores informed champion promotion; they also partly motivated M2 and informed the robustness gates and the later tolerance benchmark. The final model gives identical weights on a rerun, valid weights on a 123-month dataset and identical weights under Python 3.13, and three teammates reproduced every headline metric on their own laptops.

3 The submitted model

The forecast and risk layers add size-group structure to pooled estimates, and the portfolio layer adds an adaptive exposure overlay (Figure 2).

Size groups in forecasts and risk; adaptive exposure overlay

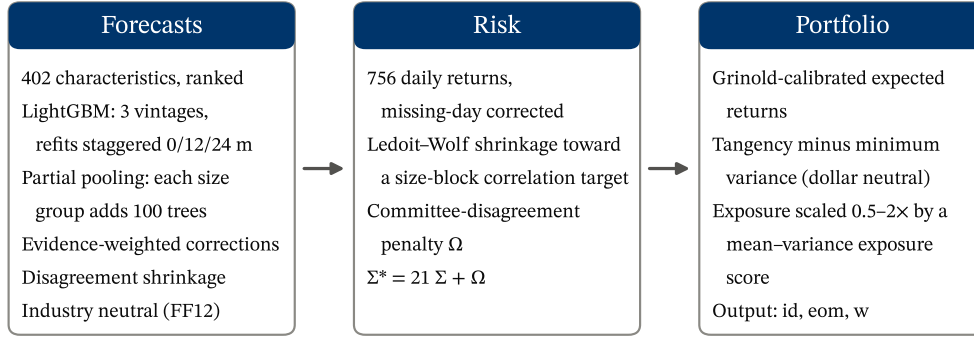


Figure 2. The submitted model.

Forecasts. As in Gu, Kelly and Xiu (2020), each characteristic is ranked within its month to $(0, 1]$, with missing values kept. The target is $\arctan(z_{i,t})$, where $z_{i,t}$ is the next-month excess return standardized within the month. The learner is LightGBM (Ke et al., 2017; 200 trees, learning rate 0.05, 255 leaves, at least 2,000 observations per leaf) on an expanding window. Refit timing alone moved the design Sharpe ratio by about ± 0.2 , so we average three vintages offset by 0, 12 and 24 months. Let b be the pooled forecast and c the corrected forecast, in which the base model grows 100 extra trees on each size group's rows at each refit. We use $b + \lambda_g(c - b)$, with $\lambda_g = \frac{1}{4} + \frac{1}{2} \text{IC}_c^+ / (\text{IC}_c^+ + \text{IC}_b^+)$, where IC^+ is the positive part of each forecast's trailing 60-month out-of-sample rank IC in that group; $\lambda_g = \frac{1}{2}$ with a shorter record or when both are zero. The vintages are z-scored, averaged and shrunk by $\tau^2 / (\tau^2 + s_i^2 / K)$, where s_i^2 is their disagreement and τ^2 the signal variance. Finally, we remove Fama–French 12-industry averages, because industry bets carried half the design-period risk but earned a Sharpe ratio of only 0.45.

Risk. Let S_t be the sample covariance of the trailing 756 daily returns of stocks with at least 250 observed days, with missing days set to zero and each series rescaled to its observed-day variance (zero-filling alone understated thin stocks' variance by up to 2.4 times). We shrink it toward a size-block target:

$$\hat{\Sigma}_t = (1 - \delta_t) S_t + \delta_t F_t, \quad (F_t)_{ij} = \bar{\rho}_{g(i)g(j),t} \sqrt{S_{ii,t} S_{jj,t}} \quad (i \neq j), \quad (F_t)_{ii} = S_{ii,t}, \quad (1)$$

where $\bar{\rho}_{ab,t}$ is the average correlation between stocks in size groups a and b (small, large, mega, other). The intensity δ_t follows Ledoit and Wolf (2004) with their single average correlation replaced by $\bar{\rho}_{g(i)g(j),t}$; with one group, (1) is exactly their estimator. The model then adds a committee-disagreement penalty,

$$\Sigma_t^* = 21 \hat{\Sigma}_t + \Omega_t, \quad (2)$$

where 21 converts daily into monthly variance and Ω_t is the variance of the mean of six committee members (three vintages, with and without corrections), with scores in return units ($\times \overline{IC}_t \sigma_t$, below). The members share most of their data, so Ω_t is a penalty, not a validated error covariance.

Portfolio. Scores become expected returns by Grinold and Kahn’s (2000) rule, $\alpha_{i,t} = \overline{IC}_t \sigma_t \text{score}_{i,t}$, where $\overline{IC}_t$ is the base forecast’s trailing rank IC (up to 60 months, shrunk toward 0.05 with a 12-month weight, floored at zero) and σ_t the trailing 12-month mean of cross-sectional return dispersion. Weights w_t are tangency minus global minimum variance under Σ_t^* (dollar neutral, unit gross exposure), scaled by

$$s_t = \text{clip}\left(\frac{1}{2} \frac{q_t}{\text{median}(q_{t-60}, \dots, q_{t-1})} + \frac{1}{2}, 0.5, 2\right), \quad q_t = \frac{\alpha_t^\top w_t}{w_t^\top \Sigma_t^* w_t}, \quad (3)$$

where the mean-variance exposure score q_t is expected return over variance ($s_t = 1$ with under 12 months of history).

4 How we arrived at the model

We built a strong conventional backbone first (M0–M3) and then tested original components on top of it (M4–M8). M4–M6 sought significant improvement; from M7 onward, we prioritized creativity subject to prespecified performance-tolerance gates. Because each component is additive, each can be tested against its absence and against random versions of itself (Table 1). In our replication of a public entry, scored on the test period, mean-variance weights doubled the Sharpe ratio of rank weights on the same forecasts (3.60 against 1.78), so we treated risk as seriously as prediction. M1 taught us to add rather than replace: swapping the shrunk covariance for a factor risk model improved design results but failed on test, because 25 factors underestimated the risk of 2,500 stocks; we then added the no-fading and risk-honesty gates. Industry neutralization (M2) removed unpaid industry bets, and staggered vintages (M3) removed refit-timing luck; together they built the reference model. LLM-designed peer-network signals raised the design Sharpe ratio slightly (6.81–6.97 against 6.70), but random peers did as well (M4). M5 replaced our own choice of signals with a walk-forward rule that keeps a signal only if it beats its placebo; placebos again matched the small gain. Blending in size specialists (M6) added up to 0.04, not significantly; specialists alone, like group-specific corrections (M7), weakened the mega-cap leg, so we moved size structure into the covariance.

Table 1. Nine versions: what each tried and what survived. Design Sharpe ratios use 1975–1989, test Sharpe ratios 1990–2023. Gates compare each version with its parent.

Version	Idea	Design SR	Test SR	Outcome
M0	LightGBM, Ledoit–Wolf, tangency – GMV	5.11	3.37	Baseline
M1	Factor risk model replaces the covariance	5.96	2.70	Rejected: risk underestimated
M2/M2.1	Industry-neutral forecasts (M2.1: bug fix)	6.04	3.43	Promoted
M3	Vintage ensemble, missing-day correction	6.70	3.58	Promoted ($t = 2.52$)
M4	LLM-designed peer-network signals	6.81–6.97	—	Small gain; random peers did as well
M5	Same signals, walk-forward placebo selection	6.81	—	Small gain; placebo did as well
M6	Separate size specialists	6.74	—	Blends +0.03–0.04 (n.s.); alone 6.43
M7	Partial pooling + evidence weights	6.61	3.63	Failed breadth by 0.025; override test
M8	M7 + size-block risk, uncertainty, timing	6.68	3.58	Final: all gates; random blocks 30/30

Source: version ledgers, pre-registrations and gate files in the team repository.

Why M8. M8 replaces M7 because it repairs M7’s weakness. M7’s group-specific trees weakened the mega-cap leg on design data, so M7 was not promoted. Moving size structure into the covariance raised that leg from 1.62 to 2.65, improved every size leg and both halves, and cut risk-forecast bias by 6%, although the overall design gain was small (6.68 against 6.61). In choosing among variants, we weighed creativity against performance and stability. The full variant tests all three new ideas at a small, insignificant cost: the size-block covariance alone scored 6.77 against 6.68 ($t = 0.54$), with similar stability across halves, size legs and risk forecasts. On test data, the mega-cap leg earned 1.61 against 1.28 and maximum drawdown fell from 9.8% to 8.4%; M7’s higher test Sharpe ratio (3.63 against 3.58) is not statistically significant under a Jobson–Korkie test ($p = 0.45$; return correlation 0.97).

5 Controlled evidence

The size-block covariance has the strongest support, from three comparisons that isolate different things (Table 2; the test comparison is in Section 4). On design data, adding it to M7 raises the Sharpe ratio from 6.61 to 6.77 ($t = 4.72$). Inside the final model, it beats all 30 versions with randomly shuffled size groups ($p = 0.03$, the minimum for 30 draws; Figure 3), which isolates the grouping, given the M8 architecture and design data.

Table 2. Component evidence: each comparison, its reference variant and what it isolates. Effects are changes in Sharpe ratio unless stated. $p = (\text{draws not beaten} + 1)/31$, one-sided, unadjusted.

Component	Comparison	Period	Effect on Sharpe	Control	Status
Industry-neutral forecasts	M2 vs M0	Design	+0.93	—	Supported; test +0.06 ($t = 1.35$)
Staggered vintages	M3 vE vs M2.1	Design	+0.54	Design ablation; seed placebo ineffective	Supported; M3 bundle test $t = 2.52$
Missing-day correction	M3 vZED vs vED	Design	+0.14	—	Supported only in combination
Disagreement shrinkage	M3 vED vs vE	Design	−0.02	—	No evidence of benefit
Group trees	M7 vH100 vs M3	Design	−0.24 ($t = -4.74$)	Random-group control not run	Lowers Sharpe; weaker mega leg
Evidence weights	M7 vHE vs vH100	Design	+0.15	—	Recovers part of the loss
	M8 vUKT vs random weights	Design	6.68 vs median 6.68	Random weights in $[\frac{1}{4}, \frac{3}{4}]$: 15/30	No demonstrated advantage over random weighting
Size-block covariance	M8 vK vs M7	Design	+0.15 ($t = 4.72$)	—	Supported
	M8 vUKT vs shuffled groups	Design	6.68 vs median 6.52	Random blocks: 30/30, $p = 0.03$	Supported, within M8
	M8 vs M7, mega leg	Test	1.61 vs 1.28	—	Consistent; not isolated
Disagreement penalty Ω	M8 vUK vs vK	Design	−0.05 ($t = -2.57$)	Wrong stocks, within vUK: 0/30	Underperforms its controls
Exposure scaling	M8 vUKT vs vUK	Design	−0.03	Shuffled scales, within vUKT: 27/30, $p = 0.13$	Inconclusive

Source: gate files and placebo audits in the team repository; design data 1975–1989 unless marked Test.

The remaining component tests provide mixed or negative evidence. Group-specific forecast trees cost 0.24 on design data and weakened the mega-cap leg; their random-group control was not run, because M7 failed its gates first. Evidence weights recovered 0.15 in M7; the later audit within M8 found no demonstrated advantage over random weighting. The disagreement penalty (2) costs 0.05 and underperforms all 30 versions placed on the wrong stocks. Exposure scaling (3) beats 27 of 30 shuffled versions ($p = 0.13$) but not constant exposure, so it is inconclusive; on design data it raised the worst 12-month return by about 3 points and lowered downside volatility, at the cost of a slightly deeper maximum drawdown. The forecast corrections were inherited from M7 and never tested for removal; the creativity-first rule kept Ω and exposure scaling. We did not re-tune after the audit.

Multiple testing. To account for the six versions scored on the test period, we compute the deflated Sharpe ratio (Bailey and López de Prado, 2014) over them (M0–M3, M7, M8; M2.1 repeats M2), in monthly units with raw kurtosis, treating trials as independent. From the spread of their Sharpe ratios, the expected null maximum is 0.45 annualized, and the deflated Sharpe ratio is near one.

6 Test-period performance

The final model earns the same test Sharpe ratio as the backbone, M3 (3.581 against 3.580), with the smallest drawdown of any version on this single path. The supported contributions are more accurate risk forecasts and the size-block finding, not Sharpe ratio. Predictability has decayed (Figure 4, Table 3). The rolling 36-month Sharpe ratio falls from about 10 in the late 1990s to about 1 after 2020, as Hellum et al. (2025) also report; the observed gross Sharpe ratio in 2020–2023 is 1.29. These are past results, not a forecast; the 10% volatility scaling is retrospective and needs about 6.6 times leverage. Average monthly turnover is 1.13 in unit-gross weights (target weights, ignoring drift), and much of the return comes from small caps. Net

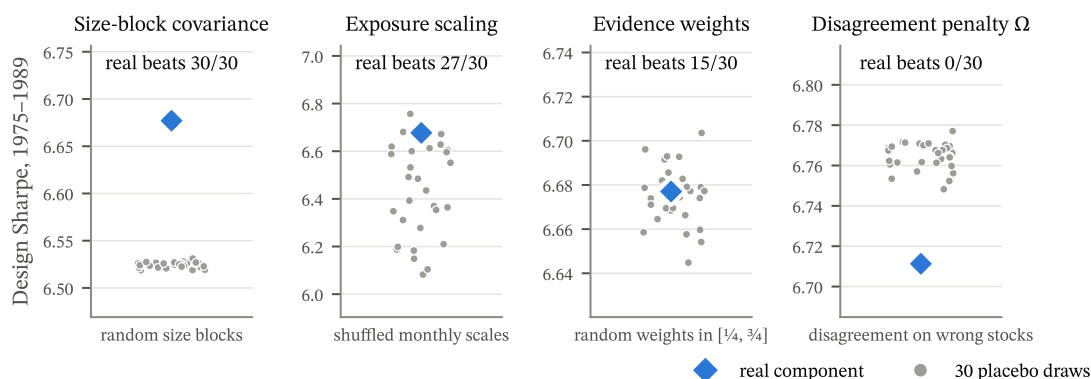


Figure 3. Only the size-block grouping clearly beats its controls. Each panel compares the real component (diamond) with 30 random versions (dots), all else fixed, on separate scales. Panels use the final model; the penalty panel uses it without exposure scaling (vUK).

Source: authors' calculations on design data, 1975–1989, with 30 placebo draws per component.

performance has not been established.

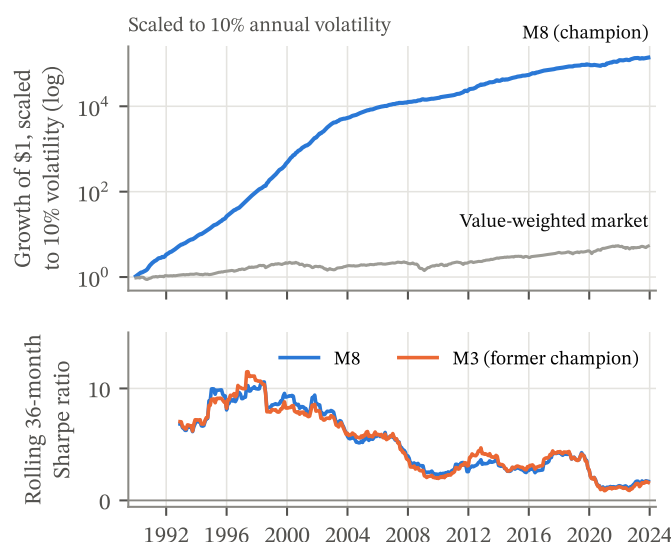


Figure 4. Strong over the full period, but decayed toward a Sharpe ratio of about 1. Top: growth of \$1 at 10% volatility, before costs. Bottom: rolling 36-month Sharpe ratio; M3's similar path shows the decay is common to the approach.

Source: returns January 1990–December 2023 on portfolios formed December 1989–November 2023; hypothetical, gross of costs.

Table 3. Test-period metrics of the final model and its two predecessors. Test period, 408 months; beta against the value-weighted CTF universe.

Metric	M8 (final)	M7	M3
<i>Full test period</i>			
Sharpe ratio	3.581	3.628	3.580
Max drawdown	−8.43%	−9.79%	−10.74%
Worst 12 months	−5.10%	−6.33%	−7.68%
Beta (CAPM)	0.09	0.09	0.09
Leverage to 10% vol	6.59×	7.25×	7.20×
<i>Sharpe ratio by decade</i>			
1990s	7.53	7.65	7.69
2000s	3.66	3.76	3.70
2010s	2.87	3.31	3.29
2020–23	1.29	1.27	1.14

7 Limitations and disclosures

Repeated test-period evaluation creates potential research-selection bias. The six-trial deflated Sharpe ratio is a limited diagnostic and does not correct for the full development process. Our first version follows the structure of Geertsema's public entry (ranked LightGBM, arctan target, 756-day covariance, tangency minus minimum variance); we wrote our own code, and every version from M2 on is our own design. An AI assistant (Claude) helped with code, literature review and documentation and proposed the M4–M5 signals; no language model runs inside the submitted model, and the team made every modeling decision.

References. Bailey and López de Prado (2014), The deflated Sharpe ratio, *JPM*. Geertsema, LightGBM with penalised VCV, CTF leaderboard entry. Grinold and Kahn (2000), *Active Portfolio Management*. Gu, Kelly and Xiu (2020), Empirical asset pricing via machine learning, *RFS*. Hellum, Jensen, Kelly and Pedersen (2025), The power of the common task framework. Ke et al. (2017), LightGBM, *NeurIPS*. Ledoit and Wolf (2004), Honey, I shrunk the sample covariance matrix, *JPM*.