Axiom Tilt: Reinforcement Learning and Large Language Models for Portfolio Optimization

Kavin Ravi[1] and Dylan Massaro[1]

[1]Chapman University , kravi@chapman.edu, dylanmassaro13@gmail.com

September 30, 2026

Abstract

We study whether two recent advances, domain-adapted language models and deep reinforcement learning (RL), improve weekly long-only equity allocation after realistic trading costs. We build a two-stage pipeline on a point-in-time S&P 500 universe (2002–2024, roughly 600 names with fundamentals coverage): a LightGBM LambdaRank ranker, augmented with 79 principal components of FinBERT embeddings of 227,000 SEC filings (10-K, 10-Q, 8-K) after continued masked-language-model pretraining, selects thirty candidates each Friday; a PPO agent then allocates weights among them. The agent's action is a log-tilt on a deterministic softmax-of-scores rule (Score-Prop), so that a zero action reproduces the rule exactly. We evaluate on sixteen expanding-window walks (806 out-of-sample weeks, 2009–2024) at 5 bps per dollar traded. Three findings emerge. First, text features improve every allocator's after-cost Sharpe ratio (by 0.08 to 0.16) relative to the same pipeline without text, although the ranker's own out-of-sample rank IC is unchanged (0.004 either way) and only the PPO pair is individually significant. Second, the RL allocator does *not* improve on the deterministic rule. For the reported seed the Sharpe difference is $+0.018$ over 2009–2024 (block-bootstrap 95% interval $[-0.03, +0.07]$); across three seeds it is $+0.001 \pm 0.015$, it is -0.005 once the configuration-selection year 2009 is excluded, and on an untouched 2025 holdout all three seeds lose to the rule. The one positive estimate is the seed whose 2009 bets paid off, selected on 2009. Third, none of the equity-only strategies beats a buy-and-hold S&P 500 ETF on a risk-adjusted basis: they are $\beta \approx 1.5$ portfolios whose CAPM alphas (5–7% per year) are not statistically significant, equal weight on the whole universe matches their Sharpe ratio, and a 60/40 blend with a Treasury ETF beats all of them. We document the model-selection protocol, seed variance, and survivorship caveats in detail, and argue that the null RL result is the informative one for practitioners.

1 Introduction

Recent breakthroughs in machine learning have generated renewed interest in applying learned systems to financial decision-making. Two developments are particularly relevant. First, large language models built on the Transformer architecture [1] have demonstrated strong capability in representing natural-language text, and financial institutions and vendors have begun building domain-specific models [2, 3]. Second, reinforcement learning has matured as a framework for sequential decision-making, with notable successes in robotics, game-playing, and recommendation systems, and a growing literature applying it to trading and portfolio construction [4]. Whether these advances translate into measurable improvements for portfolio management after costs remains an open empirical question.

The motivation for combining the two is direct. Public companies disclose substantial information through written documents (annual reports, quarterly filings, and material-event disclosures) that conventional stock-selection methods either ignore or reduce to coarse summaries, despite evidence that the textual content of these filings carries return-relevant information [5, 6]. A domain-adapted language model offers a way to extract continuous semantic representations from these filings at scale. At the same time, constructing a portfolio is fundamentally sequential: today's holdings affect tomorrow's trading costs and risk exposures. Reinforcement learning is a natural framework for such path-dependent problems.

Both components face well-known obstacles in financial settings. Language models trained on general corpora encode the register of news and conversational text more than that of regulatory disclosures, and naive transfer yields representations dominated by document structure rather than economic content. Reinforcement learning in finance is constrained by limited data (a few hundred decisions per year at weekly rebalancing), non-stationarity across regimes, and the difficulty of separating policy improvement from sampling noise [4].

This paper presents a study that takes both constraints seriously. We construct a two-stage pipeline. A supervised cross-sectional ranker, augmented with text features from a continued-pretrained FinBERT encoder [3], identifies the top thirty candidates each week from a point-in-time S&P 500 universe.[1] A PPO-based agent [7] then allocates weights across those thirty names under long-only and full-investment constraints and a linear transaction cost. The decomposition lets each subsystem use the form of supervision best suited to it: the ranker has abundant cross-sectional labels, while the allocator faces a small, path-dependent decision problem.

The key design choice on the RL side, and the source of the project's name, is a *tilt* parameterization: the agent's action is a log-space perturbation of a deterministic rule that weights candidates by a softmax of their ranker scores (Score-Prop). A zero action reproduces Score-Prop exactly, so the agent starts from a sensible policy and learns deviations from it (Section 5.4). This makes the comparison between the agent and Score-Prop a clean test of the marginal value of RL: the two share the universe, features, costs and constraints, and differ only in whether weights are learned.

We evaluate on sixteen expanding-window walks covering 2009–2024 (806 out-of-sample weeks) against Score-Prop, an equal-weight top-30 rule (EW-Top30), a minimum-variance rule, a 60/40 blend with a Treasury ETF, buy-and-hold cap-weighted and equal-weight S&P 500 ETFs, equal weight on the full universe, and text-free ablations of the three learned or score-based portfolios. Our conclusions are more cautious than the project set out to reach. Text features help downstream portfolio outcomes consistently. The RL allocator ties the deterministic rule within noise; its apparent edge belongs to one seed and one year, the year that was also used to choose its configuration, and it disappears with two more seeds and a 2025 holdout. None of the equity-only portfolios delivers a statistically significant risk-adjusted improvement over the index. We regard the transparent reporting of these null results, alongside the protocol that produced them, as the paper's main contribution.

2 Related work

Reinforcement learning for portfolio construction. Direct policy search for trading dates back to Moody and Saffell [8], who optimized differential Sharpe ratios by recurrent reinforcement learning. Deep variants followed: Deng et al. [9] combined representation learning with direct

[1] Because fundamentals coverage from our provider is incomplete, only about 600 of the roughly 900 companies that were index members during the sample are included; see Section 4. The universe is therefore survivorship-*aware* but not survivorship-*free*.

reinforcement, Jiang et al. [10] introduced the ensemble-of-identical-evaluators architecture for cryptocurrency portfolios, and Liu et al. [11] packaged common environments and algorithms in the FinRL library. Cong et al. [12] construct portfolios end to end with a transformer policy and interpret it with polynomial surrogates. Bai et al. [4] survey the field and catalogue the recurring failure modes: non-stationarity, small effective sample sizes, and the difficulty of distinguishing learning from luck. Our two-stage design restricts RL to a thirty-asset allocation problem anchored on a deterministic rule precisely to reduce exposure to those failure modes, and our evaluation focuses on whether the learned policy is distinguishable from its anchor.

Text in the cross-section of returns. Tetlock [13] and Loughran and McDonald [5] established that the tone of news and 10-K filings predicts returns and that general-purpose sentiment dictionaries misclassify financial language. Ke et al. [6] learn a supervised sentiment score from news with a sparse topic model. Neural encoders specialized to finance include FinBERT variants trained on financial communications [3, 14] and, at larger scale, BloombergGPT [2]. We follow the encoder route because a 110M-parameter model can be continued-pretrained on 227,000 filings on a single GPU and embedded at scale.

Machine learning in the cross-section and benchmark discipline. Gu et al. [15] compare tree ensembles and neural networks for return prediction and find gradient-boosted trees competitive; we use LightGBM [16] with a LambdaRank objective [17]. DeMiguel et al. [18] show that naive $1/N$ diversification is hard to beat out of sample, which motivates our equal-weight baseline. Bailey et al. [19] and Harvey et al. [20] document how configuration search inflates backtest performance; Section 6 reports our own configuration count and treats the selection year separately for that reason.

3 Problem definition

We study weekly long-only portfolio allocation over a point-in-time universe of large-capitalization US equities reconstructed from historical S&P 500 membership. At each Friday close t, a policy produces target weights $w_t \in \mathbb{R}^K$ over a candidate set of $K = 30$ names. Two constraints are enforced: no shorting ($w_{i,t} \geq 0$) and full investment ($\sum_i w_{i,t} = 1$). An earlier version of the system also capped individual weights at 10%; the cap was removed for the final configuration because it lacked a principled justification and consistently hurt the learned policy (Section 6). Trading from the previous portfolio to the new one incurs a linear cost of $c = 5$ bps per dollar of turnover.

Let $r_{i,t}$ be the realized return of asset i over week t (Friday close to Friday close; see Section 8 on execution timing) and let $\tilde{w}_{t-1}$ be the previous week's weights after drift and after pro-rata redistribution of any names that left the candidate set. The net portfolio return is

$$R_t \;=\; \sum_{i=1}^{K} w_{i,t}\, r_{i,t} \;-\; \frac{c}{10^4} \sum_{i=1}^{K} \lvert w_{i,t} - \tilde{w}_{i,t-1} \rvert. \tag{1}$$

The evaluation objective is the annualized after-cost Sharpe ratio of R_t with a zero risk-free rate, $\mathrm{SR} = \bar{R} \cdot 52 \,/\, (\hat{\sigma}_R \sqrt{52})$ computed on the compounded annualized return (Appendix A); all Sharpe and Sortino figures in this paper use this convention.[2]

[2]The 3-month Treasury bill averaged roughly 1% per year over 2009–2024, so excess-return Sharpe ratios would be about 0.03 lower for the equity-only strategies.

Solving this problem end to end with one learner over the full universe is poorly conditioned: the action space scales with the universe, the weekly reward is weak relative to noise, and any training window contains at most a few hundred decisions. We therefore decompose the policy into two stages with disjoint roles. A *cross-sectional ranker* scores every eligible asset each week and selects the top $K = 30$; an *allocator* then chooses weights over those thirty names given the ranker scores, a small set of structured features, macro state and the current portfolio.

This decomposition motivates two research questions.

RQ1 Does augmenting the ranker with domain-adapted text features derived from SEC filings improve *downstream portfolio outcomes* (after-cost return, Sharpe ratio, drawdown) relative to the same pipeline trained on structured features alone?

RQ2 Given a fixed top-K candidate set, does an RL allocator achieve higher after-cost Sharpe than the deterministic rule (Score-Prop) that it is anchored on?

RQ1 is phrased in terms of portfolio outcomes because, as Section 7.3 shows, the text features do not measurably change the ranker's cross-sectional accuracy even though they change the portfolios built on it. We answer both questions with a walk-forward backtest over 2009–2024, and we add a third, unplanned comparison, against the index itself, because it turned out to change the interpretation of both.

4 Data

The data layer joins four sources into a single point-in-time panel that feeds both the FinBERT continued pretraining and the downstream ranker and allocator.

Universe. Historical S&P 500 membership intervals are reconstructed from public constituent-change records and mapped to SEC central index keys and CRSP permanent identifiers. A stock is eligible on a date only if it was a constituent on that date.

Prices and returns. Daily prices and returns come from the Center for Research in Security Prices (CRSP) through Wharton Research Data Services. CRSP is the academic standard for US equity prices and is the only widely available source that includes delisting returns, the final realized return of a stock that exits through bankruptcy, merger or index removal. Without them a backtest understates downside in any year with delistings [21].

Fundamentals. Fundamentals come from Sharadar's SF1 dataset via Nasdaq Data Link. We use only as-reported quarterly and annual values (ARQ, ARY) rather than restated ones, and join on each statement's filing date so that no fundamental enters the panel before it was public. Of the roughly 900 companies that were index members during the sample, only about 600 have usable Sharadar coverage; the remaining ∼300, mostly older delisted names, are excluded. This is the source of the survivorship caveat in Section 8. Because the gap is concentrated in early years, the eligible set grows from roughly 300 names in 2002 to about 500 in later years, although the index itself always had 500.

Filings. Text comes from the SEC's EDGAR archive: 10-K annual reports, 10-Q quarterly filings and 8-K material-event disclosures for every company in the universe. The raw corpus contains 226,919 filings across 614 companies, about 746 GB of SGML envelopes. A multi-stage cleaning

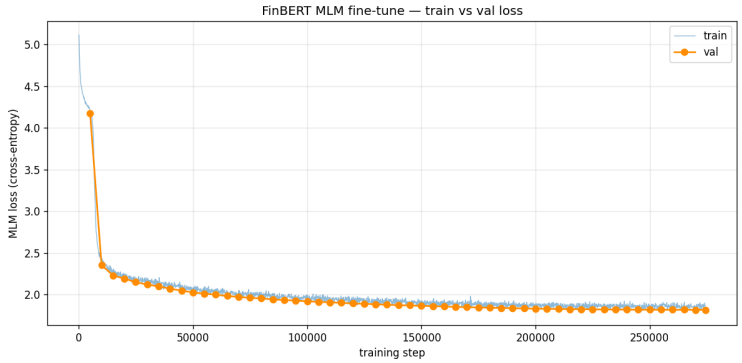


Figure 1: Masked-language-model loss during continued pretraining of FinBERT on the EDGAR corpus (one epoch, 274,387 steps).

pipeline strips HTML and XBRL markup and base64-encoded attachments, then applies a sliding-window content-density filter to remove residual table dumps and namespace noise; 226,915 cleaned filings (53 GB of text) survive. This corpus is used both for continued pretraining (Section 5.1) and for document embedding.

Macro. Four FRED series enter as features: the 10-year and 3-month Treasury yields, the 10-year minus 2-year spread [22], and the VIX index.

Benchmark. For the index comparison in Section 7 we use SPY daily closes, aligned to the same Friday rebalance dates as the strategies.

5 Method

5.1 Text encoder: continued pretraining of FinBERT

The encoder is FinBERT [3], a BERT-base model [23] (110M parameters) pretrained on financial communications. We continue masked-language-model pretraining on the cleaned EDGAR corpus to move the representation from general financial text toward the regulatory register of SEC filings. Filings are split into 512-token chunks, giving 17,560,754 training and 924,251 validation chunks (about 9.5 billion token positions). We train for one epoch (274,387 optimizer steps, batch size 64, bf16) with full-parameter updates rather than a parameter-efficient method; at this model size a full fine-tune is inexpensive on a single consumer GPU. Training was stable (Figure 1): final training loss 1.86, validation loss 1.82, validation perplexity 6.16.

Section 7.4 repeats the walk-1 text branch with the base checkpoint to measure what this step contributes.

Document and stock-day embeddings. Each cleaned filing is embedded by running the encoder over 512-token windows (stride 64) and mean-pooling the `[CLS]` vectors, giving one 768-dimensional vector per filing. For each stock and calendar day we form an exponentially weighted mean of the filing vectors published in the trailing 30 days (14-day half-life), forward-filled between filings, yielding 2,998,336 stock-day vectors over 579 stocks. Three auxiliary scalars accompany the vector: *text novelty* (one minus the cosine similarity between today's vector and the vector seven calendar days earlier), *days since last filing*, and the *number of filings in the last seven days*.

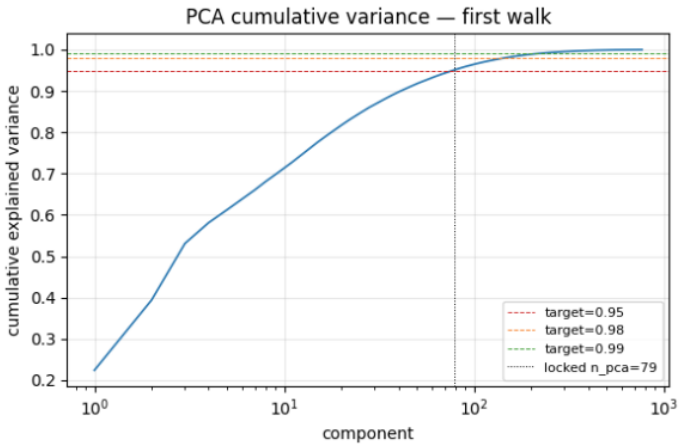


Figure 2: Cumulative explained variance of the stock-day embeddings on the first walk's training window (2002–2007). Dashed lines mark 95%, 98% and 99%; the dotted line is the locked $n_{\text{pca}} = 79$.

5.2 Dimensionality reduction

Treating each of the 768 embedding dimensions as a separate tree feature invites the ranker to latch onto a handful of them. We therefore project the stock-day vectors with principal component analysis fitted on the training window of each walk only. On the first walk, 95%, 98% and 99% cumulative variance are reached at 78, 145 and 211 components (Figure 2); the 99% target would leave a 211-dimensional vector, not a meaningful reduction from 768. We locked $n_{\text{pca}} = 79$ (95.1% of variance on walk 1) for all walks, refitting the basis at every walk boundary.

5.3 Cross-sectional ranker

The final feature set per stock-week is: 79 PCA text components, 3 auxiliary text scalars, 4 macro series, 104 Sharadar fundamentals and 8 CRSP price/volume fields, about 190 features in total. Each Friday the ranker scores roughly 300–500 eligible stocks (one row per stock, not a single 95,000-dimensional vector).

The ranker is a LightGBM [16] model with the LambdaRank objective [17]. Each Friday is one query group. The label is the 32-quantile bucket of the stock's five-day forward return in excess of the cross-sectional mean on that Friday. A regression objective on the raw excess return was also tried and was clearly worse on both validation NDCG@30 (0.176 vs 0.233 on walk 1) and test rank IC; the pairwise loss is more robust to the heavy tails of weekly returns. Hyperparameters (number of leaves, learning rate, minimum leaf size, feature and bagging fractions, ℓ_2 penalty) were tuned with 15 Optuna [24] trials on the first walk's validation year (2008) and then frozen for walks 2–16; the number of boosting rounds is chosen by early stopping on validation NDCG@30 in every walk (Appendix B).

Out-of-sample ranking quality is modest. The mean per-Friday Spearman rank information coefficient (IC) across the sixteen test years is 0.0042; it is positive in ten walks and negative in six (Figure 3), with a mean top-minus-bottom decile spread of 26 bps per week, most of it from 2009. The text-free ranker is statistically indistinguishable on the same metrics (Table 5). The downstream portfolio results in Section 7 show that this weak signal is nonetheless enough to select a top-30 set with market-level risk-adjusted returns and much higher raw returns.

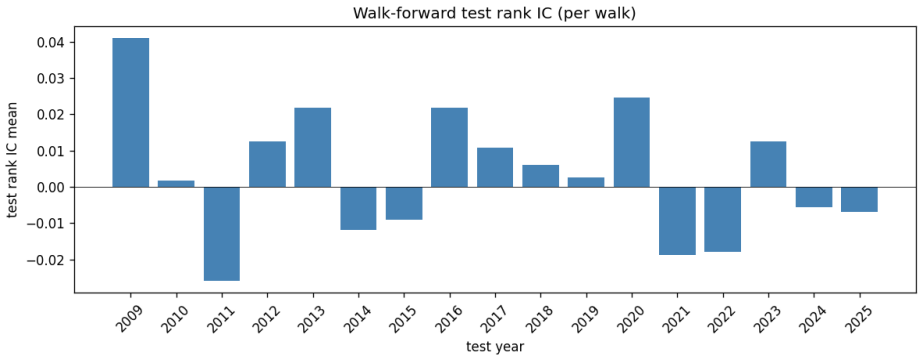


Figure 3: Out-of-sample mean rank IC of the text-augmented ranker per walk (test year on the horizontal axis).

5.4 Allocator: PPO with a tilt on the deterministic rule

Deterministic anchor. Given ranker scores $s_t \in \mathbb{R}^K$ for the top-K set, the Score-Prop rule sets

$$w_t^{\mathrm{SP}} = \mathrm{softmax}(s_t), \tag{2}$$

optionally followed by a water-filling cap at a maximum weight (inactive in the final configuration). Note that this is a softmax of scores, not weights proportional to scores.

Tilt action. The agent outputs $a_t \in [-5, 5]^K$ and the portfolio is

$$w_t = \mathrm{softmax}(\log w_t^{\mathrm{SP}} + a_t), \tag{3}$$

so that $a_t = 0$ reproduces Score-Prop exactly and the policy learns log-space deviations from it. Without this anchor, no configuration in our first search round matched Score-Prop (Section 6); with it, the policy starts at the rule's performance and can only be judged on what it adds.

State. The 214-dimensional observation contains the current (drifted) weights (K); the ranker scores z-scored within the date (K); five structured features per name (payout ratio, dividend cash flow, daily low, SG&A, retained earnings; the top-five by ranker gain on walk 1), z-scored within the date ($5K$); three macro series (VIX, 10-year yield, 10y–2y spread); and last week's realized portfolio return. Observations are standardized with running statistics that are frozen at evaluation time.

Reward. Let $e_t = w_t^\top r_t - \frac{1}{K}\mathbf{1}^\top r_t$ be the week's return in excess of the equal-weight top-K portfolio and cost_t the transaction cost in (1). The reward is a rolling Sharpe proxy over the last $W = 16$ weeks of the episode,

$$\rho_t = \frac{\mathrm{mean}(e_{t-W+1:t})}{\mathrm{std}(e_{t-W+1:t}) + 10^{-6}} - \mathrm{cost}_t, \tag{4}$$

falling back to $e_t - \mathrm{cost}_t$ for the first two steps of an episode.

Training. Episodes are 104 consecutive Fridays (two years) starting at a random Friday in the walk's training window; names that leave the top-K have their weight redistributed pro rata to entrants, and the resulting forced trade is charged at the next step. We use PPO [7] as implemented

in Stable-Baselines3 [25] with four parallel environments, one million environment steps per walk, a [128, 64] MLP policy, learning rate 10^{-4}, clip range 0.15, target KL 0.03, entropy coefficient 0.02, GAE $\lambda = 0.95$ and $\gamma = 0.99$ (full list in Appendix B). The policy checkpoint with the best deterministic return on the walk's validation year is kept. The text-free ablation uses identical environment and PPO settings.

6 Experimental protocol

Walk-forward design. We use sixteen expanding-window walks. Walk n trains the PCA, ranker and allocator on 2002 through 2006+n, validates on 2007+n, and tests on 2008+n; walk 1 therefore tests on 2009 and walk 16 on 2024. Per-walk test weeks are concatenated into one 806-week out-of-sample series per strategy (2 January 2009 to 20 December 2024) and all metrics are computed on that series. Ranker hyperparameters and the PCA dimension are fixed after walk 1. A seventeenth walk (test year 2025, 49 weeks) is used only in Section 7.6, as a holdout that no selection step touched.

Strategies. All strategies share the same weekly top-30 candidate set (from the text-augmented ranker unless marked *NoText*), the same 5 bps linear cost, and the same long-only, fully-invested constraint.

- **PPO**: the tilt-parameterized RL allocator of Section 5.4.
- **Score-Prop**: the deterministic softmax-of-scores rule (2), i.e. PPO with $a_t \equiv 0$. The contrast PPO vs Score-Prop isolates the marginal value of RL.
- **EW-Top30**: equal weight on the thirty candidates [18].
- **MinVar**: minimum-variance weights from a 26-week trailing covariance with a 10% cap. Included for completeness; as discussed in Section 7 it is ill-conditioned under weekly membership churn and should not be read as a serious competitor.
- **60/40 (IEF)**: 60% EW-Top30 plus 40% in the iShares 7–10 year Treasury ETF (IEF, total return), rebalanced weekly; the bond leg is not charged a cost. An earlier version of this study used a carry-only bond proxy (the 10-year yield divided by 52, no price risk); we keep that row as "60/40 (carry proxy)" for comparison.
- **SPY** and **RSP**: buy-and-hold S&P 500 ETF and equal-weight S&P 500 ETF, Friday close to Friday close, no cost.
- **EW-Universe**: equal weight over every eligible name in our panel each Friday (gross, no cost), which isolates the ranker's selection effect from the equal-weighting effect.
- **NoText ablations**: PPO, Score-Prop and EW-Top30 rebuilt on a ranker trained without the 79 PCA components and the 3 auxiliary text scalars (108 features), with its own 15-trial Optuna search on walk 1.

Metrics. Annualized net return, annualized volatility, Sharpe and Sortino ratios (rf = 0), maximum drawdown, Calmar ratio, weekly hit rate and average weekly turnover (Appendix A).

Model selection for the allocator, and what it implies. The allocator's reward shape, episode length, action parameterization and PPO hyperparameters were chosen by an automated search [26] over roughly forty configurations (Appendix C). The search scored each configuration by its *walk-1 test-year (2009)* Sharpe relative to Score-Prop. Two facts follow and are reported here rather than hidden. First, 2009 is not a clean out-of-sample year for PPO: it was used to select the configuration, so we report all headline RL comparisons both with and without it. Second, the search revealed large seed variance: the selected configuration, retrained with three seeds on walk 1, produced excess Sharpe of -0.018, $+0.029$ and $+0.315$ relative to Score-Prop. The reported policy uses the best of these seeds. Section 7.6 reports the same configuration retrained with two further seeds on all walks, and adds 2025 as a holdout year that no selection step touched.

To measure what selecting on the validation year would have done, we re-scored every saved policy from the search on the 2008 validation window of walk 1 (Table 13 in the appendix). Among the PPO configurations the validation ranking puts the selected configuration second (excess Sharpe $+0.072$ versus -0.381 for Score-Prop in that crash year), behind a variant with a 10% per-name cap ($+0.099$) that also did well on 2009. The two SAC configurations rank first on validation and last on test, with weekly turnover above 1.4, so a validation rule would have needed a turnover screen to avoid them. Across all 39 configurations, the Spearman correlation between validation and test excess Sharpe is 0.24 ($p = 0.14$): validation-year performance is a weak guide to the following year, which is the practical reason we treat 2009 with suspicion rather than a reason to believe the selection was lucky in one particular direction.

Two other findings of the search are worth recording because they shaped the method. The rolling-Sharpe reward (4) was the single largest improvement over a raw excess-return reward, and two-year episodes were clearly better than one-year episodes, plausibly because a policy must see more than one regime within an episode to learn regime-conditional behavior. Constraints of every kind (per-name caps, a narrower action range, an information-ratio reward, a cost curriculum) made the policy worse. Training beyond one million steps did not change the result.

7 Results

7.1 Full-period performance

Table 1 reports the 806-week out-of-sample results for all strategies and benchmarks, and Figure 4 the corresponding equity curves. Three patterns are visible before any statistics.

First, among the three portfolios built on the text-augmented candidate set, PPO has the highest return and Sharpe ratio, Score-Prop is second and EW-Top30 third, and the three are within 1.2 percentage points of each other on maximum drawdown. The ordering is the one the design predicts, but the gaps are small: 0.018 in Sharpe between PPO and Score-Prop, and 0.034 between Score-Prop and EW-Top30.

Second, every text-augmented portfolio beats its text-free counterpart on return, Sharpe ratio and drawdown. The Sharpe gains are $+0.158$ (PPO), $+0.077$ (Score-Prop) and $+0.077$ (EW-Top30), and they persist on 2010–2024. This is the most consistent effect in the study.

Third, none of the equity-only portfolios beats the index on a risk-adjusted basis. SPY's Sharpe ratio of 0.856 sits between EW-Top30 and Score-Prop over 2009–2024 and above all three on 2010–2024. The candidate portfolios earn roughly twice the index's return with roughly twice its volatility and a maximum drawdown of 54% versus 32%. A 60/40 blend of the same thirty names with a Treasury ETF has the highest Sharpe ratio in the table.

Table 1: Out-of-sample performance, 2009–2024 (806 weeks), 5 bps per dollar traded, Sharpe and Sortino with rf = 0. Turnover is the average weekly one-way trade as a fraction of the book. The right-most column repeats the Sharpe ratio on 2010–2024 (757 weeks), excluding the year used for allocator configuration selection. RSP is the equal-weight S&P 500 ETF; EW-Universe is equal weight over every eligible name in our panel each Friday; 60/40 (IEF) is 60% EW-Top30 plus 40% in the 7–10 year Treasury ETF, rebalanced weekly with no cost on the bond leg.

Strategy	Ann. ret.	Ann. vol	Sharpe	Sortino	Max DD	Calmar	Turnover	$\text{Sharpe}_{10\text{–}24}$
PPO	28.87%	32.32%	0.893	1.376	−54.32%	0.531	0.353	0.790
Score-Prop	27.93%	31.90%	0.876	1.334	−54.07%	0.517	0.237	0.786
EW-Top30	25.84%	30.70%	0.842	1.251	−53.15%	0.486	0.000	0.775
PPO – NoText	24.94%	33.94%	0.735	1.110	−58.03%	0.430	0.268	0.635
Score-Prop – NoText	27.00%	33.83%	0.798	1.224	−57.09%	0.473	0.059	0.688
EW-Top30 – NoText	24.66%	32.24%	0.765	1.136	−56.13%	0.439	0.000	0.669
SPY (buy and hold)	14.87%	17.37%	0.856	1.130	−31.83%	0.467	0.000	0.845
RSP (buy and hold)	14.32%	19.28%	0.743	0.968	−36.75%	0.390	0.000	0.689
EW-Universe	16.39%	19.57%	0.838	1.071	−40.50%	0.405	0.000	0.804
60/40 (IEF)	17.18%	18.02%	0.953	1.426	−33.45%	0.514	0.000	0.903
60/40 (carry proxy)	17.26%	18.42%	0.937	1.391	−35.33%	0.489	0.000	0.877
MinVar	7.89%	37.51%	0.210	0.244	−80.70%	0.098	1.728	0.582

7.2 Is the RL allocator distinguishable from its anchor? (RQ2)

Because PPO and Score-Prop hold the same names and differ only in weights, their weekly returns can be compared as a paired series. Table 2 summarizes the comparison. Over the full period the annualized return difference is 0.9 percentage points with a paired t-statistic of 1.06; a stationary block bootstrap [27] of the two series jointly (5,000 draws, mean block length 8 weeks) gives a 95% interval on the Sharpe difference of $[-0.031, +0.066]$ and a 24% probability that the true difference is at or below zero. Excluding 2009, the year used for configuration selection, the difference collapses to +0.004 in Sharpe and 0.05 percentage points in return ($t = 0.04$).

Table 4 shows where the point estimate comes from. PPO beat Score-Prop by 30 percentage points in 2009 and 12 points in 2010, both strong recovery years in which the policy's willingness to concentrate paid off. From 2011 onward it lost to Score-Prop in eight of fourteen years, and its three largest post-2010 losses (2019–2021, −6.1, −5.4 and −2.8 points) exceed its largest post-2010 win. It also traded 49% more (0.353 versus 0.237 per week), so the cost assumption matters. Table 3 recomputes the comparison at 10 and 20 bps: the PPO advantage falls to +0.006 Sharpe at 10 bps and becomes a −0.017 deficit at 20 bps.

On this evidence alone the answer to RQ2 is already negative: the point estimate in the allocator's favour is well inside sampling noise, is driven by the year on which it was selected, and reverses at a plausible institutional cost level. Section 7.6 adds two more seeds and a holdout year and removes the point estimate as well.

7.3 Do text features help? (RQ1)

At the level of portfolio outcomes, yes, with one qualification. Every text-augmented portfolio beats its ablation on every metric in Table 1, the ordering survives the removal of 2009, and the Sharpe gains (Table 2) are two to nine times the RL gain. But only the PPO pair is individually significant at the 5% level over 2009–2024 (95% CI $[+0.005, +0.317]$); the Score-Prop and EW-Top30 pairs have intervals that include zero, and the three pairs are not independent since they share the candidate set.

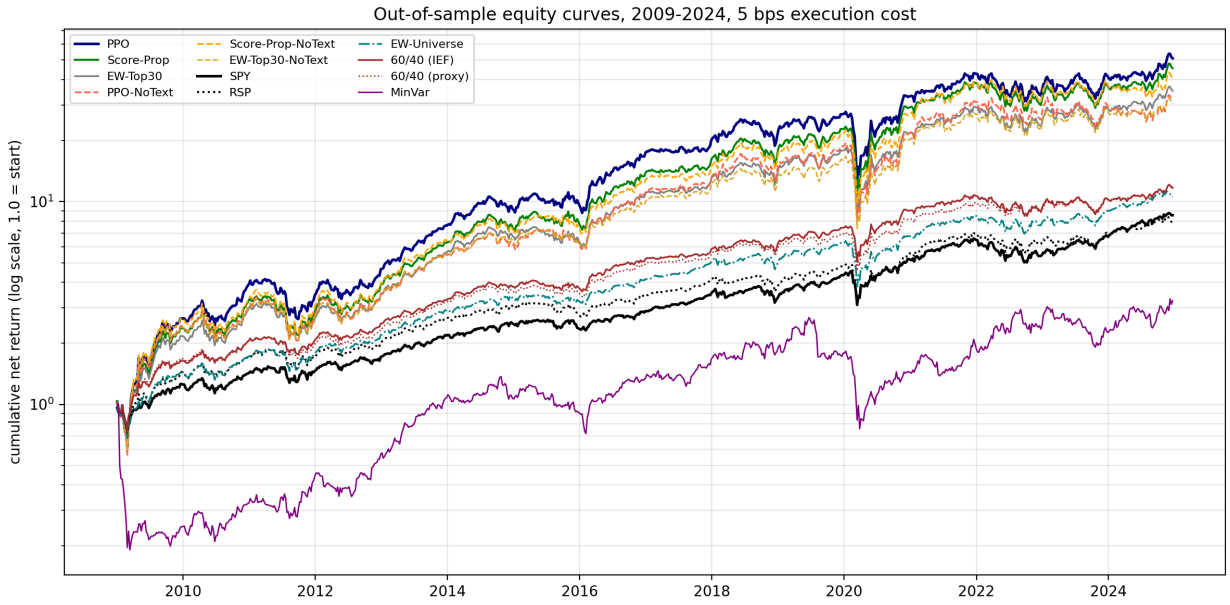


Figure 4: Cumulative net return (log scale) over the 806-week out-of-sample period at 5 bps execution cost. Dashed lines are the text-free ablations; black lines are the index ETFs.

At the level of ranking accuracy the answer is no. Table 5 scores the text-free rankers on the same sixteen test years with the same metrics used for the text-augmented ones. The mean rank IC is 0.0042 with text and 0.0036 without (paired $t = 0.45$ over walks); both are positive in ten of sixteen years; the mean decile spread is 26 versus 23 bps per week. Whatever the text features contribute to the portfolios, it is not visible in the ranker's cross-sectional ordering. The most likely channel is the composition of the top-30 set rather than its ordering: the text-augmented candidates have lower volatility (30.7% versus 32.2% for EW-Top30) and shallower drawdowns, consistent with text novelty and filing recency steering the ranker away from names in the middle of adverse disclosure events. We flag this as a hypothesis, not a finding.

7.4 Does the continued pretraining matter? (walk 1)

To isolate the contribution of the continued pretraining, we rebuilt the walk-1 text branch with the public FinBERT checkpoint and no further training: base-model embeddings of the 69,615 filings dated 2000–2009, the same 30-day stock-day aggregation, a PCA basis with the same 79 components fitted on 2002–2007, a ranker tuned with the same 15-trial Optuna budget on 2008, and the same PPO configuration and seed. Table 6 compares the three text variants on the 2009 test year.

The continued-pretrained branch beats the base branch on every column except validation NDCG, by 0.06 to 0.12 in Sharpe for the two deterministic allocators. But the no-text branch beats both text branches on EW-Top30 and Score-Prop in this year, and the base and no-text rankers have the same test IC. The two encoders also pick materially different stocks: the weekly top-30 sets of the base and continued-pretrained branches overlap with a Jaccard index of 0.45. On one year we can say only that continued pretraining changes the candidate set and did not hurt; whether it helps would need the ablation on all sixteen walks, which costs one embedding pass over the full corpus.

Table 2: Paired comparisons of weekly net returns. Δ is the first strategy minus the second. The interval is a stationary block bootstrap (Politis–Romano, mean block 8 weeks, 5,000 draws) on the difference of Sharpe ratios; $P(\leq 0)$ is the bootstrap probability that the difference is at or below zero.

Comparison	Window	Δ ann. ret.	paired t	Δ Sharpe	95% CI	$P(\leq 0)$
PPO − Score-Prop	2009–24	+0.9 pp	1.06	+0.018	[−0.031, +0.066]	0.24
PPO − Score-Prop	2010–24	+0.1 pp	0.04	+0.004	[−0.046, +0.058]	0.44
PPO − PPO-NoText	2009–24	+3.9 pp	1.18	+0.158	[+0.005, +0.317]	0.02
Score-Prop − Score-Prop-NoText	2009–24	+0.9 pp	0.07	+0.077	[−0.056, +0.225]	0.13
EW-Top30 − EW-Top30-NoText	2009–24	+1.2 pp	0.27	+0.077	[−0.049, +0.214]	0.12
PPO − PPO-NoText	2010–24	+3.3 pp	1.06	+0.155	[−0.006, +0.332]	0.03
Score-Prop − Score-Prop-NoText	2010–24	+1.8 pp	0.57	+0.098	[−0.045, +0.247]	0.09
EW-Top30 − EW-Top30-NoText	2010–24	+2.1 pp	0.80	+0.106	[−0.018, +0.245]	0.05
EW-Top30 − EW-Universe	2009–24	+9.5 pp	2.78	+0.004	[−0.227, +0.235]	0.49
PPO − SPY	2009–24	+14.0 pp	2.89	+0.037	[−0.283, +0.398]	0.42
PPO − SPY	2010–24	+8.4 pp	2.12	−0.055	[−0.400, +0.342]	0.61

Table 3: Cost sensitivity: after-cost Sharpe ratio (2009–2024) when the linear cost is 5, 10 or 20 bps per dollar traded. The equal-weight rules are unaffected because their turnover is not charged in the backtest (they are rebuilt each week from the candidate set); the Score-Prop-NoText row has low turnover because its scores are more persistent.

Strategy	5 bps	10 bps	20 bps
PPO	0.893	0.857	0.785
Score-Prop	0.876	0.851	0.802
PPO − Score-Prop	+0.018	+0.006	−0.017
PPO – NoText	0.735	0.709	0.659
Score-Prop – NoText	0.798	0.793	0.781
MinVar	0.210	0.084	−0.153

7.5 Against the index

Table 7 regresses each strategy's weekly net return on SPY. All the candidate portfolios are high-beta ($\beta \approx 1.5$) portfolios. Their annualized CAPM alphas are positive, 4.8% to 7.0% for the text-augmented variants, but the standard errors are of the same size and none reaches conventional significance. The paired bootstrap of PPO against SPY (Table 2) tells the same story: the return difference is large and significant (+14 points per year, $t = 2.9$) but the Sharpe difference is +0.04 with an interval of $[-0.28, +0.40]$, and it is negative on 2010–2024.

Two further benchmarks locate the source of the candidate portfolios' returns. Equal weight over our entire eligible universe each Friday (EW-Universe) earns 16.4% per year at Sharpe 0.838 with $\beta = 1.06$; the ranker's top-30 subset (EW-Top30) earns 9.5 points more per year ($t = 2.8$) at exactly the same Sharpe ratio (difference +0.004). The ranker therefore selects a higher-beta, higher-return slice of the universe rather than a higher-quality one. The equal-weight S&P 500 ETF (RSP) underperforms both, which suggests our panel's missing early-year constituents (Section 8) contribute part of the EW-Universe return.

Finally, the 60/40 comparison survives a real bond leg. Replacing the carry-only proxy with the 7–10 year Treasury ETF (IEF) *raises* the blend's Sharpe ratio, from 0.937 to 0.953 (0.903 on 2010–2024), because Treasuries were negatively correlated with equities for most of the sample.

Table 4: Per-year annualized net return over the out-of-sample period.

Year	PPO	Score-Prop	EW-Top30	SPY	EW-Universe	PPO − Score-Prop
2009	181.9%	151.8%	128.7%	27.9%	48.5%	+30.1
2010	50.2%	38.2%	37.7%	13.7%	19.4%	+12.0
2011	−12.3%	−11.3%	−10.6%	2.6%	3.5%	−0.9
2012	35.2%	36.1%	32.4%	17.5%	24.2%	−0.8
2013	65.7%	61.6%	61.6%	28.1%	32.4%	+4.0
2014	22.1%	20.2%	21.5%	15.1%	15.6%	+1.8
2015	10.9%	10.3%	−1.4%	−4.9%	−1.8%	+0.6
2016	73.3%	71.1%	66.1%	21.4%	27.2%	+2.2
2017	16.3%	17.0%	19.4%	23.2%	23.0%	−0.8
2018	5.2%	5.9%	4.7%	−6.1%	−3.1%	−0.6
2019	28.0%	34.1%	31.9%	30.8%	27.1%	−6.1
2020	22.9%	28.3%	22.7%	21.8%	8.9%	−5.4
2021	31.4%	34.2%	36.1%	24.9%	27.2%	−2.8
2022	−16.9%	−16.9%	−15.7%	−15.7%	−7.7%	0.0
2023	16.2%	15.7%	16.6%	22.9%	14.7%	+0.4
2024	24.4%	24.2%	23.8%	29.2%	17.5%	+0.3

The blend has half the drawdown of the equity-only portfolios at 60% of their return. The point is not specific to RL: every long-only equity allocator in this study, learned or not, has a lower Sharpe ratio than this simple multi-asset mix.

7.6 Seed variance and the 2025 holdout

Two further checks address the two weaknesses of the RQ2 comparison identified in Section 6: the reported policy is one seed, and 2009 was used to choose its configuration. We retrained the selected configuration with seeds 43 and 44 on all walks, added a seventeenth walk (train 2002–2023, validate 2024, test 2025) for all three seeds, and reran the full-period backtest. No step of the search saw 2025. Table 8 reports the results; Figure 5 shows the three equity curves against Score-Prop.

Three things follow. First, the seed that was reported is the only one that beats the anchor on 2009–2024, and its entire margin comes from 2009 and 2010: in those two years seed 42 beat Score-Prop by 30 and 12 percentage points, while seeds 43 and 44 were within one point of it in 2009 and three points behind it in 2010 (Table 9). The search selected the seed whose 2009 concentration bets paid off, and that is what the 2009 result records. Second, across seeds the full-period difference is zero to three decimal places: +0.001 in Sharpe on 2009–2024, −0.005 on 2010–2024, with seed-to-seed standard deviations of 0.009 to 0.015. Third, on the 2025 holdout all three seeds lose to Score-Prop, by 0.003 to 0.331 in Sharpe and by 0.4 to 8.7 percentage points of return; the seed-averaged difference has a bootstrap interval that excludes zero. 2025 was a year in which the candidate set itself lost money (Score-Prop −8.0%), and the policy's extra turnover (0.39 to 0.49 per week versus 0.05 for Score-Prop) bought nothing.

Including 2025, the three-seed full-period Sharpe is 0.800 ± 0.006 against 0.809 for the rule the policy is anchored on. We therefore restate the answer to RQ2: over 855 weeks and three seeds, the tilt-parameterized allocator does not improve on its deterministic anchor, and the one positive point estimate in this paper is the product of selecting on the year it was measured in.

Table 5: Out-of-sample ranking metrics per walk, text-augmented (190 features) versus text-free (108 features) ranker, evaluated on each walk's test year. IC is the mean per-Friday Spearman correlation between scores and realized five-day excess returns; the decile spread is the top-minus-bottom decile mean excess return in basis points per week.

Test year	IC (text)	IC (no text)	Decile bps (text)	Decile bps (no text)
2009	+0.041	+0.033	+200	+207
2010	+0.002	+0.002	+29	+19
2011	−0.026	−0.023	−25	−32
2012	+0.013	+0.012	+18	+16
2013	+0.022	+0.010	+50	+27
2014	−0.012	−0.016	0	−4
2015	−0.009	−0.008	−1	+3
2016	+0.022	+0.018	+49	+47
2017	+0.011	+0.007	+2	−9
2018	+0.006	+0.008	+20	+21
2019	+0.003	+0.009	0	0
2020	+0.025	+0.020	+78	+57
2021	−0.019	−0.009	+6	+2
2022	−0.018	−0.016	−23	−17
2023	+0.013	+0.015	+13	+22
2024	−0.006	−0.005	+3	+1
Mean	+0.0042	+0.0036	+26	+23

Table 6: Walk 1 (test year 2009) with three text branches. Ranker metrics are on the 2009 cross-section; portfolio metrics are annualized net of 5 bps. The PPO column for the continued-pretrained branch is the configuration-selection result and is not comparable; the EW-Top30 and Score-Prop columns are unaffected by that selection.

	Ranker			Sharpe, 2009		
Text branch	Val. NDCG@30	Test IC	Decile bps	EW-Top30	Score-Prop	PPO
Continued-pretrained FinBERT	0.220	+0.041	+200	2.02	2.38	2.70
Base FinBERT	0.225	+0.033	+151	1.96	2.26	2.34
No text	0.233	+0.033	+207	2.30	2.61	2.30

7.7 Robustness of the RL policy's own performance

Two Monte Carlo checks characterize the PPO series itself (Figure 6). A random-weights null replaces the policy's weights each Friday with a Dirichlet draw on the same thirty names; against 2,000 uniform draws the realized gross Sharpe ratio of 0.930 has $p = 0.046$, and against draws calibrated to the policy's average concentration $p = 0.004$. This shows that the policy's weights are not random, not that they beat Score-Prop, which the paired test above addresses. A stationary block bootstrap of the policy's own weekly returns (5,000 draws, mean block 8) gives a 95% interval on its Sharpe ratio of $[0.31, 1.59]$, a reminder of how wide sixteen years of weekly data leave the estimate. The policy never held more than 38% in a single name; its mean effective number of positions was 25.1 versus 27.7 for Score-Prop (Figure 8).

Table 7: Weekly CAPM regression on SPY (rf = 0), 2009–2024. Alpha is annualized; the standard error is the OLS residual standard error scaled to annual units.

Strategy	β	α (per year)	s.e. / t
PPO	1.53	7.0%	4.7% / 1.49
Score-Prop	1.52	6.2%	4.5% / 1.38
EW-Top30	1.49	4.8%	4.2% / 1.14
PPO – NoText	1.59	3.5%	5.0% / 0.70
Score-Prop – NoText	1.58	5.2%	5.0% / 1.04
EW-Top30 – NoText	1.54	3.5%	4.6% / 0.76
EW-Universe	1.06	0.8%	1.7% / 0.46
RSP	1.07	−1.2%	1.3% / −0.92

Table 8: Config 038 trained with three seeds versus its deterministic anchor. Sharpe ratios (rf = 0, 5 bps) by window; Δ is PPO minus Score-Prop. The seed-average row is an equal-weight blend of the three policies' weekly returns; its interval is a stationary block bootstrap (mean block 8, 5,000 draws) on the Sharpe difference. Seed 42 is the policy reported elsewhere in this paper and the one whose configuration was chosen on 2009.

	2009–24 (806 wk)		2010–24 (757 wk)		2025 holdout (49 wk)		2009–25 (855 wk)	
	Sharpe	Δ	Sharpe	Δ	Sharpe	Δ	Sharpe	Δ
Score-Prop	0.876	–	0.786	–	−0.318	–	0.809	–
PPO seed 42	0.893	+0.018	0.790	+0.004	−0.649	−0.331	0.803	−0.006
PPO seed 43	0.872	−0.003	0.779	−0.007	−0.321	−0.003	0.804	−0.005
PPO seed 44	0.863	−0.013	0.773	−0.013	−0.380	−0.062	0.794	−0.016
Mean ± sd	0.876 ± 0.015	+0.001	0.780 ± 0.009	−0.005	−0.450 ± 0.175	−0.132	0.800 ± 0.006	−0.009
Seed average, 95% CI		[−0.025, +0.033]		[−0.033, +0.029]		[−0.377, −0.009]		[−0.038, +0.023]

7.8 On the MinVar baseline

The minimum-variance rule lost about 80% in the first months of 2009 and traded 173% of the book per week. A 26-week trailing covariance of thirty names whose membership changes every week is ill-conditioned, and the weekly re-solve amplifies estimation noise into turnover. From 2010 onward the same rule returned 19% per year at Sharpe 0.58. We keep the row for completeness but do not draw conclusions from it.

8 Limitations

Configuration selection on the first test year. The allocator's configuration was chosen by its 2009 performance, and 2009 is the first out-of-sample year. Section 7.6 shows what this cost: the selected seed's 2009 margin is not shared by two other seeds of the same configuration. We report every RL comparison with and without 2009 and with a 2025 holdout, and the three-seed numbers are the ones we would defend. A clean design would select on the 2008 validation year only.

Three seeds are still few. The seed-to-seed standard deviation of full-period Sharpe (0.006 to 0.015) is of the same order as the effects under test. Three seeds are enough to show that the reported seed is the favourable one, not enough to bound the configuration's expected performance tightly.

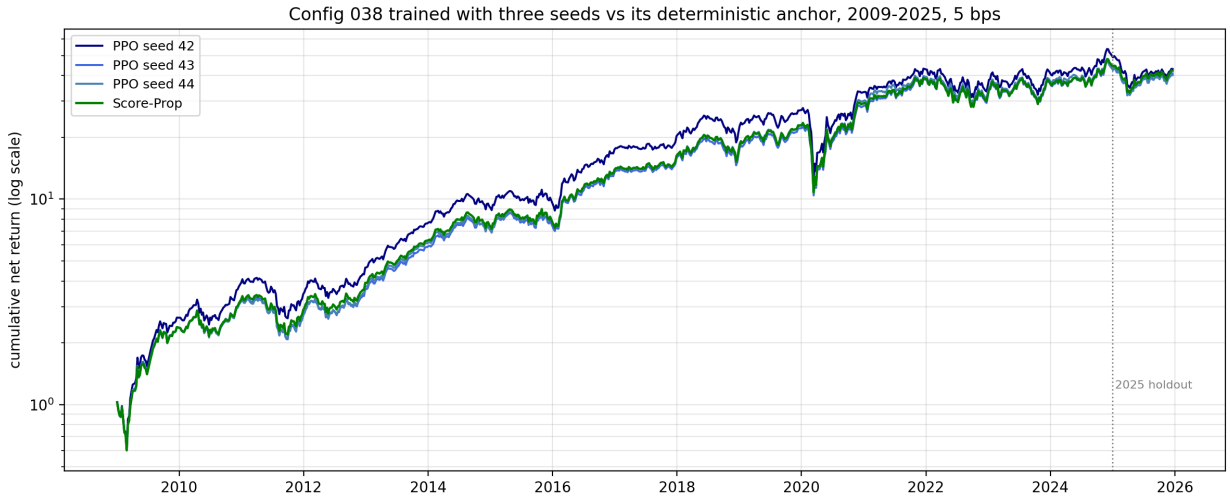
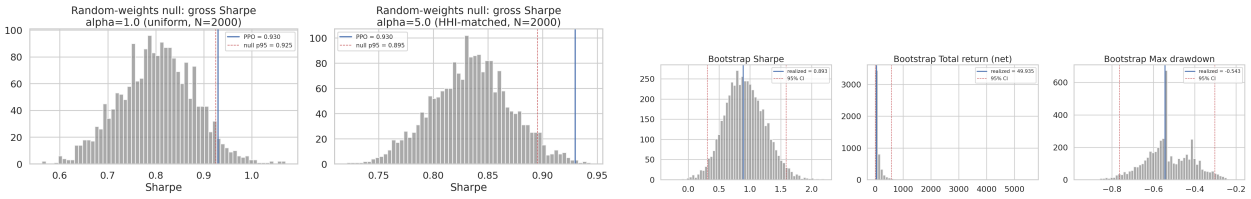


Figure 5: Cumulative net return of the selected configuration trained with three seeds, against Score-Prop, 2009–2025. The dotted line marks the start of the 2025 holdout year.



(a) Random-Dirichlet-weights null.

(b) Stationary block bootstrap of PPO returns.

Figure 6: Monte Carlo checks on the PPO out-of-sample series.

Survivorship. The universe includes delisted constituents and CRSP delisting returns, which is better than typical retail-data backtests, but about 300 of the roughly 900 historical constituents lack fundamentals coverage and are excluded, disproportionately in the early years. Returns of all candidate portfolios are likely biased upward as a result.

Execution and cost model. The label is the Friday-close-to-Friday-close five-day return, so the backtest assumes execution at Friday's close rather than the following Monday's open; a one-session lag would reduce returns slightly. The cost model is a flat 5 bps on turnover with no market impact. For a small account this is realistic; for institutional size, 10–30 bps is more plausible, and PPO's higher turnover makes it the most exposed strategy.

Continued pretraining evaluated on one walk only. The base-versus-continued-pretrained comparison (Section 7.4) covers the 2009 test year. The full-period text-feature gains in Table 1 are measured against no text, not against the base encoder, so the paper cannot say how much of them the continued pretraining is responsible for.

Weak ranking signal. The ranker's mean out-of-sample IC is 0.004 with or without text, and its decile spread is near zero in most years. The top-30 portfolios' returns are dominated by their $\beta \approx 1.5$ exposure: equal weight on the full universe has the same Sharpe ratio as equal weight on

Table 9: Per-year annualized net return of Score-Prop and the difference to each PPO seed, in percentage points. 2025 is the holdout year.

Year	Score-Prop	seed 42 − Score-Prop	seed 43 − Score-Prop	seed 44 − Score-Prop
2009	151.8%	+30.1	−0.1	−0.4
2010	38.2%	+12.0	−3.3	−2.4
2011	−11.3%	−0.9	−3.2	−3.9
2012	36.1%	−0.8	−1.6	+2.6
2013	61.6%	+4.0	+0.1	+2.8
2014	20.2%	+1.8	+3.5	−1.6
2015	10.3%	+0.6	−0.2	+4.2
2016	71.1%	+2.2	+2.9	−0.4
2017	17.0%	−0.8	+0.2	+0.4
2018	5.9%	−0.6	−1.4	+1.5
2019	34.1%	−6.1	+0.8	−2.1
2020	28.3%	−5.4	+0.1	+4.8
2021	34.2%	−2.8	+3.5	−1.0
2022	−16.9%	0.0	−0.3	+0.9
2023	15.7%	+0.4	+0.3	−1.0
2024	24.2%	+0.3	−2.1	−7.0
2025	−8.0%	−8.7	−0.4	−1.6

the ranker's top 30. We have not tested whether a beta-neutral construction on the same names retains any alpha.

Benchmarks. The MinVar rule is ill-conditioned and should be replaced by a shrinkage-based version before any conclusion is drawn from it. The EW-Universe benchmark is built on our panel, which is missing about 300 mostly delisted early-year constituents, so it inherits the same upward bias as the candidate portfolios; RSP does not, and it is 1.5 points per year lower.

Regime dependence. Sixteen years contain one financial-crisis recovery, one pandemic shock and one rate-hiking cycle. The RL policy's only material wins were in the two recovery years. All findings are conditional on this sample.

9 Conclusion

We built and evaluated an end-to-end pipeline that combines a continued-pretrained financial language model, a LambdaRank stock ranker, and a PPO allocator anchored on a deterministic softmax-of-scores rule, on a point-in-time S&P 500 universe with CRSP delisting returns, as-reported fundamentals and 227,000 SEC filings, evaluated over sixteen expanding walks and 806 out-of-sample weeks.

On RQ1 the answer is a qualified yes: adding FinBERT-derived text features to the ranker improved after-cost return, Sharpe ratio and drawdown for all three allocation rules built on it, with Sharpe gains of 0.08 to 0.16 that survive the removal of the configuration selection year, although only the PPO pair is individually significant. The gain does not come from better ranking: the text-free ranker has the same out-of-sample IC. It comes from a different, lower-volatility candidate mix. On the one walk where we rebuilt the text branch with the base checkpoint, continued pretraining

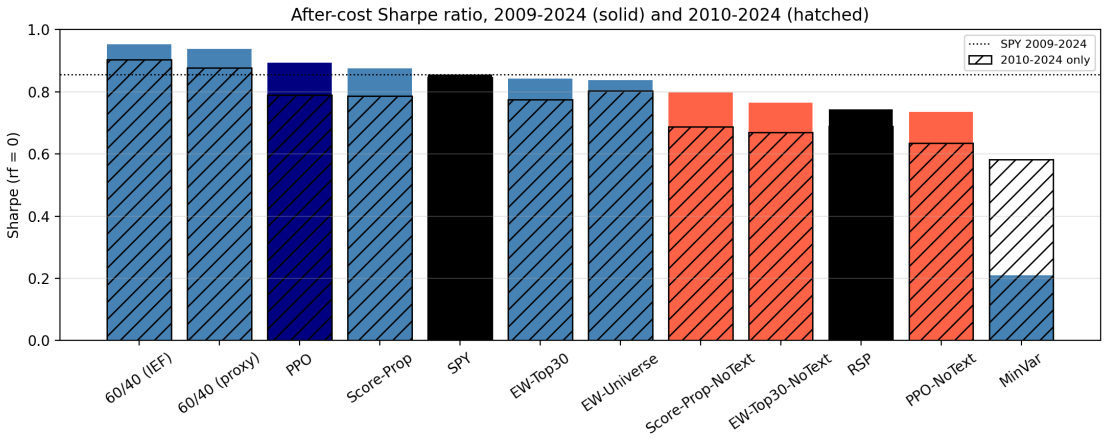


Figure 7: After-cost Sharpe ratio by strategy, 2009–2024 (solid) and 2010–2024 (hatched). The dotted line is SPY over 2009–2024.

changed the candidate set and improved the deterministic allocators by 0.06 to 0.12 in Sharpe, but a no-text ranker did better still in that year.

On RQ2 the answer is no. For the reported seed the full-period Sharpe difference of +0.018 has a bootstrap interval that comfortably includes zero and shrinks to +0.004 without 2009. Across three seeds the difference is +0.001, and on a 2025 holdout that no selection step touched, all three seeds lose to the anchor. The reported seed is the one whose 2009 concentration bets paid off, and 2009 is the year it was selected on. The allocator trades 34% to 38% of the book each week for this.

The comparisons the original study omitted turned out to matter most: none of the equity-only portfolios beats a buy-and-hold index ETF on a risk-adjusted basis, equal weight on the whole universe has the same Sharpe ratio as equal weight on the ranker's top 30, and a 60/40 blend with a Treasury ETF beats every equity-only row. The candidate portfolios are leveraged-beta portfolios with positive but insignificant alpha.

What remains open is narrower than where the project started: a shrinkage-based minimum-variance baseline, the continued-pretraining ablation on all walks rather than one, and a selection protocol that never scores a test year. The tilt parameterization itself, which lets an RL agent start from a competent rule and be judged only on its deviations, is the part of the design we would keep.

Data and code availability

The full pipeline (ingestion, cleaning, continued pretraining, embedding, feature assembly, ranker, allocator, backtest, and robustness checks) is open source at `https://github.com/kavinravi/axiom_tilt`. CRSP is licensed through WRDS; Sharadar SF1 requires a Nasdaq Data Link subscription; EDGAR and FRED are free. Trained artifacts (the continued-pretrained encoder, per-walk PCA bases, ranker checkpoints, and PPO policies) are too large for the repository but are reproducible from the code given the licensed inputs.

Acknowledgements

CRSP daily price and delisting data were obtained through Wharton Research Data Services under Chapman University's subscription. Fundamentals are from Sharadar's SF1 dataset, licensed by the authors through Nasdaq Data Link. Filings were downloaded from the SEC's EDGAR archive and macroeconomic series from the Federal Reserve Bank of St. Louis (FRED); both are public. SPY, RSP and IEF prices are from Yahoo Finance. All training and evaluation, including the continued pretraining of FinBERT, ran on a single consumer workstation owned by the first author (one NVIDIA RTX 5090, one AMD Ryzen 9 9950X); no grant or institutional compute was used. The work began as a project in a Chapman University course taught by Tim Frenzel, whom we thank for his continued guidance and support throughout the project.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[3] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. FinBERT: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

[4] Yahui Bai, Yuhe Gao, Runzhe Wan, Sheng Zhang, and Rui Song. A review of reinforcement learning in financial applications. *arXiv preprint arXiv:2411.12746*, 2024.

[5] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

[6] Zheng Tracy Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting returns with text data. Technical Report w26186, National Bureau of Economic Research, 2019.

[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[8] John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

[9] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2016.

[10] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.

[11] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. In *Deep RL Workshop, NeurIPS 2020*, 2020.

[12] Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. *SSRN working paper 3554486*, 2021.

[13] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[14] Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[15] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[16] 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*, volume 30, 2017.

[17] Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

[18] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

[19] David H. Bailey, Jonathan Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014.

[20] Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

[21] Tyler Shumway. The delisting bias in CRSP data. *The Journal of Finance*, 52(1):327–340, 1997.

[22] Arturo Estrella and Frederic S. Mishkin. Predicting U.S. recessions: Financial variables as leading indicators. *The Review of Economics and Statistics*, 80(1):45–61, 1998.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[24] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[25] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

[26] Andrej Karpathy. autoresearch: AI agents running research on single-GPU nanochat training automatically. https://github.com/karpathy/autoresearch, 2026.

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

A Metric definitions

For a weekly net return series $R_1, \ldots, R_n$ with n weeks: total return $T = \prod_t (1 + R_t) - 1$; annualized return $A = (1 + T)^{52/n} - 1$; annualized volatility $\sigma = \sqrt{52}\,\mathrm{sd}(R)$; Sharpe $= A/\sigma$; Sortino $= A/\sigma_-$ with σ_- the annualized standard deviation of negative weeks; maximum drawdown $= \min_t(E_t/\max_{s\le t} E_s - 1)$ for the equity curve $E_t = \prod_{s\le t}(1 + R_s)$; Calmar $= A/|\mathrm{MDD}|$; hit rate = fraction of weeks with $R_t > 0$; turnover = mean of $\sum_i |w_{i,t} - \tilde{w}_{i,t-1}|$. The risk-free rate is set to zero throughout.

B Hyperparameters

Table 10: Ranker (LightGBM LambdaRank). Search ranges are the Optuna space on walk 1 (15 TPE trials, median pruner, objective validation NDCG@30); the chosen values are frozen for walks 2–16. Boosting rounds are chosen by early stopping (patience 50) on each walk.

Parameter	Search range	With-text / no-text
`num_leaves`	15–127	17 / 57
`learning_rate`	0.01–0.1 (log)	0.037 / 0.089
`min_data_in_leaf`	20–200	86 / 152
`feature_fraction`	0.6–1.0	0.74 / 0.84
`bagging_fraction`	0.6–1.0	0.84 / 0.66
`lambda_l2`	0–5	3.45 / 0.78
`n_estimators` (max)	2000	early stopped (246 / 31 on walk 1)
Label buckets	–	32
Features	–	190 / 108

Software. Versions are taken from the saved artifacts. The continued pretraining of FinBERT used Transformers 4.57 and PyTorch 2.9; the ranker used LightGBM 4.6 with Optuna 4.8 and scikit-learn 1.7; the PPO policies reported in Tables 1–7 were trained with Stable-Baselines3 2.8 on PyTorch 2.9 (CUDA 12.8), and the seed-43, seed-44 and walk-17 policies of Section 7.6 and the ablation of Section 7.4 with the same Stable-Baselines3 on PyTorch 2.13 (CUDA 13.0). Python 3.12 throughout. Hardware: one NVIDIA RTX 5090 (bf16) and one 16-core AMD Ryzen 9 9950X.

C Allocator configuration search

Table 11: Allocator (PPO, Stable-Baselines3 2.8). Identical for the with-text and no-text runs.

Parameter	Value
Action	$[-5, 5]^{30}$, log-tilt on Score-Prop (Eq. 3), tilt scale 1.0
Observation	214-dim, VecNormalize (clip 10)
Reward	rolling Sharpe of excess return vs equal weight, window 16, minus cost
Episode length	104 weeks, random start in training window
Max weight	1.0 (no cap)
Cost	5 bps per dollar traded
Policy network	MLP [128, 64], tanh
Learning rate	10^{-4}
`n_steps` / batch	2048 / 64
Epochs per update	5
Clip range / target KL	0.15 / 0.03
Entropy / value coefficients	0.02 / 0.5
γ / GAE λ	0.99 / 0.95
Max gradient norm	0.5
Parallel environments	4
Total steps per walk	10^6
Checkpoint selection	best deterministic validation-year return, evaluated every 10^4 steps
Seed	42

Table 12: Selected configurations from the two search rounds, scored on walk 1 (test year 2009). "Excess" is PPO Sharpe minus Score-Prop Sharpe (2.380 on this walk). The full log of all configurations is `experiments/results.tsv` in the repository.

Id	Change relative to previous best	Excess Sharpe	Turnover
001	PPO, raw excess-return reward, 1M steps	−0.859	0.63
003	rolling-Sharpe reward, window 8	−0.308	0.15
008	window 16	−0.188	0.37
023	entropy coefficient 0.02	−0.153	0.43
028	tilt anchor on Score-Prop, 52-week episodes	−0.002	0.38
038	104-week episodes (selected)	+0.315	0.46
042 / 043	same as 038, seeds 43 / 44	+0.029 / −0.018	0.21 / 0.23
044	156-week episodes	+0.165	0.46
045	038 with 5M steps	+0.318	0.46
037	038 with 10% cap	−0.136	0.38
034 / 035	SAC without / with tilt	−0.009 / −0.854	1.26 / 1.53

Table 13: Re-scoring the saved search policies on the walk-1 *validation* year (2008) instead of the test year (2009). Excess Sharpe is PPO minus Score-Prop on the same dates (Score-Prop Sharpe: −0.381 in 2008, +2.380 in 2009). Top ten by validation excess Sharpe plus the three seed replicates of the selected configuration; full table in `paper/tables/reselect_on_validation.csv`.

Id	Configuration	Val. excess	Val. turnover	Test excess
035	SAC + tilt	+0.259	1.48	−0.854
002	SAC, excess-return reward	+0.243	1.56	−0.828
046	038 + 10% cap	+0.099	0.43	+0.360
038	**selected**	**+0.072**	0.46	**+0.315**
045	038, 5M steps	+0.071	0.46	+0.318
013	no tilt, 10 epochs	+0.038	0.20	−0.323
040	038, action range 2	+0.036	0.49	−0.068
039	038, Sharpe window 32	+0.034	0.41	−0.078
044	156-week episodes	+0.028	0.46	+0.165
043	038, seed 44	+0.026	0.25	−0.018
042	038, seed 43	−0.016	0.22	+0.029

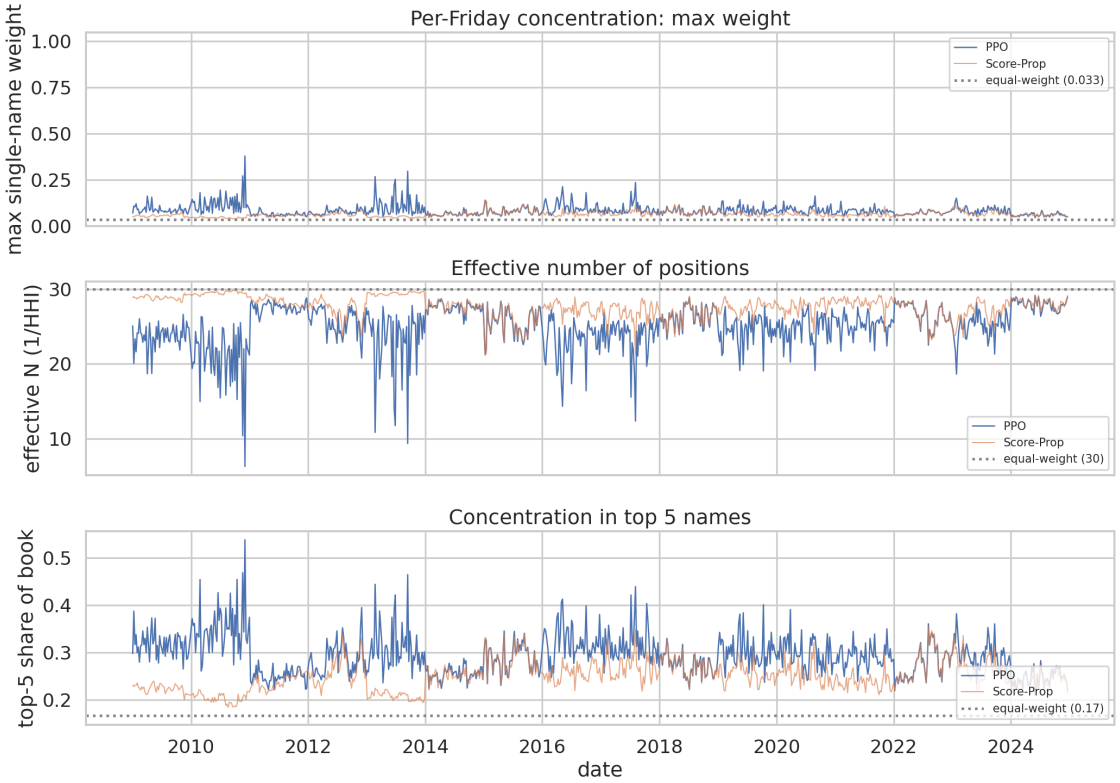


Figure 8: Per-Friday concentration of the PPO and Score-Prop portfolios over the out-of-sample period: maximum single-name weight, effective number of positions (1/HHI), and top-5 share.