# PLAN BEFORE YOU TRADE: INFERENCE-TIME OPTIMIZATION FOR RL TRADING AGENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning agents for portfolio management are typically trained and deployed as static policies, with no mechanism for policy adjustments using price forecasts at inference time. We propose FinPILOT (**Fin**ancial **P**lugin **I**nference-time **L**earning for **O**ptimal **T**rading), a plugin inference-time policy optimization framework inspired by Model Predictive Control (MPC). Under the price-taker assumption, portfolio actions do not affect future market prices, allowing a multi-step price forecaster to provide imagined market trajectories without iterative action-conditioned rollouts. At each decision step, FinPILOT uses the forecaster's predicted price trajectory to construct an allocation-based imagined return objective, and optimizes the pre-trained policy at inference-time before executing one step of the trade. Our framework is compatible with any pre-trained policy and adapts the policy using the forecaster's predictions without modifying the original RL training pipeline. Evaluated with five pre-trained policies (PPO, SAC, A2C, TD3, DDPG) across DJ30 equities and foreign exchange (FX) over three rolling held-out test years (30 settings in total), FinPILOT improves Total Return in 28 of 30 settings and Sharpe ratio in 27 of 30, with gains spanning both stochastic and deterministic policies. Direct single- and multi-period portfolio optimizers supplied with the same forecasts do not reproduce these aggregate gains under our evaluation protocol. Further, using controlled forecasts at calibrated quality levels, we show that both the magnitude and consistency of FinPILOT's performance increase as forecast error is reduced.

## 1 INTRODUCTION

Reinforcement learning for financial trading (RLFT) has emerged as a principled framework for sequential portfolio decision-making, with recent benchmarks reporting competitive performance across a range of market conditions (Liu et al., 2022a; Sun et al., 2023). At the same time, machine learning for trading has developed along two largely separate lines: forecasting-based and RL-based methods.

The forecasting-based line of work treats prediction as the central problem, using increasingly sophisticated models for short-horizon returns or price movements, ranging from the linear-to-tree spectrum systematized by Gu et al. (2020) to LSTM and Transformer-based movement predictors (Feng et al., 2019; Yoo et al., 2021). Out-of-sample financial predictability is typically weak, with $R^2$ often below 1% at monthly horizons (Gu et al., 2020). Moreover, in this paradigm, the mechanism translating forecasts into portfolio allocations is designed separately from the predictor.

The RL-based line of work instead learns the trading decision directly. Building on the deep RL formulation of Jiang et al. (2017), subsequent methods add risk-aware objectives (Wang et al., 2021), augmented state representations carrying price-prediction signals (Ye et al., 2020), and standardized benchmarking platforms (Sun et al., 2023). Forecasts can therefore be incorporated during policy training, but this does not provide a general way to attach an independently trained forecaster to an already pre-trained policy at inference time.

A related line of model-based RL performs planning through learned world models, including Dreamer (Hafner et al., 2020; 2025), MuZero (Schrittwieser et al., 2020), and TD-MPC2 (Hansen et al., 2024). Most directly related to our setting, Deb et al. (2026) perform inference-time gradient-based policy adaptation through a learned differentiable world model on continuous-control offline



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

RL benchmarks. This demonstrates inference-time policy optimization in that setting, but leaves open how to connect such adaptation to an independently trained financial forecaster.

The world-model perspective suggests a natural route. Model Predictive Control (MPC) repeatedly optimizes over a finite future horizon, executes the first resulting action, and replans after observing the next state (García et al., 1989; Mayne et al., 2000). Translating this paradigm to RLFT raises two key questions: **(i)** What should an imagined world model contain in financial markets? **(ii)** Given a forecasted future, should one optimize portfolio actions directly, as in classical financial MPC (Boyd et al., 2017), or adapt the parameters of a pre-trained policy?

For **(i)**, price-taking portfolio management has a useful structural decomposition. Under the price-taker assumption, the agent's allocation does not affect subsequent market prices, so the uncertain market component of an imagined rollout can be supplied by an exogenous multi-step forecaster rather than an action-conditioned transition model. Conditional on a forecasted market path and the policy's allocations, portfolio wealth, transaction costs, weight drift, and rewards follow exactly from known portfolio mechanics. Thus, imagined rollouts decompose into an independently trained market forecaster and closed-form transaction-cost-aware portfolio accounting, without requiring a learned reward model.

For **(ii)**, we adapt the pre-trained policy itself rather than optimizing an action sequence. At each real decision step, the forecaster provides an $H$-step future trajectory, against which we optimize the actor parameters using imagined portfolio returns and a downside-risk penalty. Only one real action is then executed before the procedure is repeated, and the adapted actor parameters are carried forward to the next decision. The forecaster is trained independently, and the RL policy has no access to its outputs during the original training procedure. We call this framework `FinPILOT` (***Fin****ancial* ***P****lugin* ***I****nference-time* ***L****earning for* ***O****ptimal* ***T****rading*).

We evaluate `FinPILOT` across five RL algorithms, two asset classes (DJ30 and FX), and three rolling held-out test years. Across 30 settings, `FinPILOT` improves Total Return in 28 and Sharpe ratio in 27, and direct portfolio optimizers supplied with the same forecasts do not reproduce these aggregate gains. Controlled forecast-skill experiments further show that the magnitude and consistency of the gains increase as forecast error is reduced.

**Contributions.** We summarize our contributions as follows:

1. **A modular forecast-to-policy interface.** We introduce `FinPILOT`, which uses an independently trained multi-step forecaster to adapt a pre-trained RL trading actor at inference time. Exploiting the price-taker structure of portfolio management, imagined rollouts combine exogenous forecasted market paths with exact transaction-cost-aware portfolio accounting, so no learned action-conditioned transition or reward model is needed and the original RL training pipeline is left unchanged.
2. **Broad paired evaluation.** Across five RL algorithms, two asset classes, and three rolling held-out years, `FinPILOT` improves the frozen pre-trained policy in 28 of 30 comparisons on Total Return and 27 of 30 on Sharpe ratio, while direct portfolio optimizers given the same forecasts do not reproduce these gains under our protocol.
3. **A controlled forecast-skill protocol.** We rescale the errors of the forecasters used in our main experiments to an exact target $R^2_\star$ and show that both the magnitude and consistency of `FinPILOT`'s Sharpe gains increase with forecast skill, while the original low-skill forecasts already yield consistent Sharpe improvements.

## 2 Problem Formulation

Following the RLFT literature and the TradeMaster benchmarking framework (Liu et al., 2022b; Sun et al., 2023; Zhang et al., 2024), we formalize portfolio management over a universe of $N$ assets as a Markov Decision Process (MDP) defined by the 5-tuple $(S, A, \mathbb{P}, R, \gamma)$, where $S$ is the state space, $A$ is the action space, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state-transition function determined by market dynamics, $R : S \times A \to R$ is the scalar reward function, and $\gamma \in [0, 1)$ is the discount factor. At each trading step $t$ (e.g., a day), the agent allocates its capital across the assets for step $t + 1$.

**State Space.** At each time step $t$, the state $s_t \in S$ encodes all information available to the agent. For each asset $i \in \{1, ..., N\}$, we compute a vector of 11 temporal features from historical daily

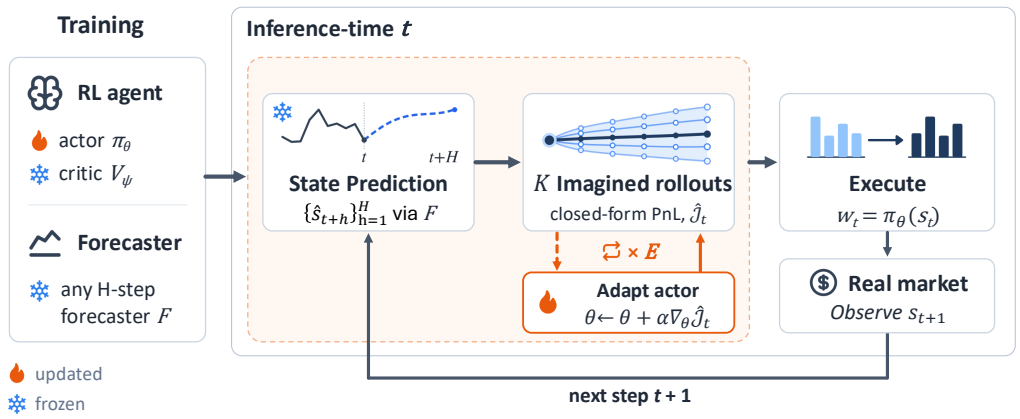



Figure 1: Overview of FinPILOT. **Training:** an RL agent is pre-trained without any forecaster, and an $H$-step forecaster $F$ is trained separately. **Inference (at time $t$):** $K$ imagined trajectories are sampled once around $F$'s predicted trajectory and reused across $E$ epochs, during which only the actor is updated by gradient ascent on the risk-penalized objective $\widehat{\mathcal{J}}_t$ (closed-form PnL with a terminal bootstrap from the frozen critic). The adapted actor executes $w_t = \pi_\theta(s_t)$, observes $s_{t+1}$, and repeats. Mini-plots are schematic.

prices (open, high, low, close, adjusted-close), following the feature set of TradeMaster. We refer the reader to Appendix B for the full feature definitions.

**Action Space.** The action $a_t \in \mathbb{R}^{N+1}$ is mapped via softmax to a portfolio weight vector $w_t = (w_{0,t}, w_{1,t}, \ldots, w_{N,t})$ on the simplex, where $w_{0,t}$ is the cash weight and $w_{i,t}$ $(i \geq 1)$ is the fraction allocated to asset $i$. Short selling and leverage are precluded by construction since softmax outputs satisfy $w_{i,t} \geq 0$ and $\sum_i w_{i,t} = 1$. Cash earns zero return and provides an uninvested reserve. Because the action space is continuous, any actor-critic algorithm that outputs continuous actions can be used to learn the policy.

**State Transition.** The transition function $\mathbb{P}$ is included in the MDP 5-tuple for completeness, but the agent does not learn a model of $\mathbb{P}$ during training: all five algorithms we consider are model-free and update their actor and critic directly from observed transitions. Under the standard price-taker assumption used in the RLFT literature (Liu et al., 2022b; Sun et al., 2023), the market-state transition is independent of the agent's portfolio allocation, i.e., $\mathbb{P}(s_{t+1} \mid s_t, a_t) = \mathbb{P}(s_{t+1} \mid s_t)$. This approximation is reasonable when the agent's trades are small relative to available market liquidity, such that their price impact is negligible. It may break down for sufficiently large or aggressive trades, for which the agent's own trading activity can affect subsequent market prices. At inference time, FinPILOT uses a separately trained price forecaster $F$ to predict an $H$-step future market-state trajectory for imagined rollouts during inference-time adaptation (Section 3). Under the price-taker assumption, this does not require an action-conditioned transition model.

**Reward Function.** The training reward is the step-wise portfolio Profit and Loss (PnL). We index each reward by the decision step that generates it. At time $t$, after selecting the target portfolio $w_t$, the agent incurs a transaction cost $\delta_t = c_{\text{tx}} U_t \left\| w_t - w_t^- \right\|_1$, where $U_t$ is the portfolio value immediately before rebalancing and $w_t^-$ denotes the post-price-movement, pre-rebalancing portfolio weights. Given the realized next-period price relatives $x_{t+1}$, portfolio value evolves as $U_{t+1} = (U_t - \delta_t)\, w_t^\top x_{t+1}$. The reward associated with the decision at time $t$ is therefore $r_t = U_{t+1} - U_t$. The imagined rollouts in FinPILOT use the same portfolio accounting, with realized price relatives replaced by forecast-derived ones; the full multi-step formulation is given in Section 3.

The pre-trained agent maximizes expected discounted cumulative reward,

$$\pi_\theta^* = \arg\max_{\pi_\theta} \mathbb{E}_{\pi_\theta} \left[ \sum_{k=0}^{T-t-1} \gamma^k r_{t+k} \right].$$

No explicit volatility, drawdown, or downside-risk penalty is used during RL training. Risk sensitivity is introduced only during inference-time adaptation.

**Algorithm 1** FinPILOT: inference-time policy adaptation

**Require:** Pre-trained actor $\pi_\theta$, frozen critic $\mathcal{V}_\psi$, forecaster $F$, horizon $H$, particles $K$, epochs $E$, step size $\alpha$, discount $\gamma$, risk weight $\lambda$, noise scale $\sigma$

**for** $t = 1, \ldots, T$ **do**
  Observe $s_t$, $w_t^-$, and portfolio value $U_t$
  $\{\hat{s}_{t+h}\}_{h=1}^{H} \leftarrow F(s_{\leq t})$
  Sample and cache $K$ perturbed trajectories $\{\tilde{s}_{t:t+H}^{(k)}\}_{k=1}^{K}$
  **for** $e = 1, \ldots, E$ **do**
    Compute $\{J_t^{(k)}(\theta)\}_{k=1}^{K}$ using equation 1
    Compute the Monte Carlo estimate $\widehat{\mathcal{J}}_t$ of equation 2
    $\theta \leftarrow \theta + \alpha \nabla_\theta \widehat{\mathcal{J}}_t$
  Execute $w_t = \pi_\theta(s_t)$

## 3 FinPILOT

We assume a reinforcement learning agent has been pre-trained on historical market data using a model-free actor-critic algorithm with continuous action outputs. We denote the actor parameters by $\theta$ and critic parameters by $\psi$, with $\pi_\theta$ the actor and $\mathcal{V}_\psi$ the critic. During training, the agent learns from observed market transitions with no access to forecasted future prices.

### 3.1 Inference-Time Policy Adaptation

At inference time, we introduce an independently trained $H$-step forecaster $F$, absent during RL training, that provides imagined future market trajectories for policy adaptation. FinPILOT is forecaster-agnostic: $F$ may be deterministic or probabilistic, is trained only on data preceding the evaluation period, and remains frozen during inference. Algorithm 1 summarizes the inference-time adaptation procedure; a fully expanded version is provided in Algorithm 2 in the appendix. Figure 1 illustrates the overall framework.

**Forecast uncertainty.** FinPILOT does not prescribe a particular representation of forecast uncertainty. It only requires a collection of imagined future trajectories over the planning horizon. A probabilistic forecaster can provide these trajectories directly by sampling from its predictive distribution. For deterministic point forecasters, such as the XGBoost and Theta + VAR(1)-ridge models we use, we construct $K$ forecast-error scenarios around the predicted trajectory.

Specifically, for each forecast horizon $h$, we estimate a cross-asset residual covariance matrix $\Sigma_h^{\text{val}}$ from forecast errors on the validation split and freeze it before test-time evaluation. At decision time $t$, particle $k$ uses $\varepsilon_{t,h}^{(k)} \sim \mathcal{N}\left(0,\, \sigma^2 \Sigma_h^{\text{val}}\right)$ to perturb the corresponding point forecast and form an imagined trajectory $\{\tilde{s}_{t+h}^{(k)}\}_{h=1}^{H}$. The $K$ sampled trajectories are generated once at each decision step and held fixed across all $E$ inference-time adaptation epochs. Gradients are not propagated through either the forecaster or the trajectory-construction procedure.

**Closed-form imagined return.** A useful property of portfolio management is that, conditional on a market trajectory and portfolio allocations, the resulting portfolio reward is available in closed form. Thus, FinPILOT does not require a learned reward model: the forecaster supplies the imagined market trajectory, while portfolio PnL follows deterministically from the corresponding price relatives, allocations, and transaction costs.

Consider a single imagined trajectory, omitting the particle index for clarity. At the beginning of the rollout, $U_t$ and $w_t^-$ are the current real portfolio value and post-price-movement, pre-rebalancing portfolio weights, respectively. We set $\tilde{s}_t = s_t$. At imagined step $t + h$, for $h = 0, \ldots, H - 1$, the actor produces the target allocation $w_{t+h} = \pi_\theta(\tilde{s}_{t+h})$. The transaction cost incurred when rebalancing to $w_{t+h}$ is $\delta_{t+h} = c_{\text{tx}} U_{t+h} \left\| w_{t+h} - w_{t+h}^- \right\|_1$, where $U_{t+h}$ is the portfolio value immediately before rebalancing. We apply $c_{\text{tx}} = 10$ bps to the full $\ell_1$ turnover over the complete portfolio, including cash. Thus, a complete transfer from one portfolio sleeve to another has turnover 2 and incurs a 20 bp cost.

Given the forecast-derived next-period price relatives $\tilde{x}_{t+h+1}$, imagined portfolio value evolves as

$$U_{t+h+1} = (U_{t+h} - \delta_{t+h})\, w_{t+h}^{\top} \tilde{x}_{t+h+1}.$$

Consistent with the real-environment reward defined in Section 2, we index the imagined reward by the decision step that generates it: $\hat{r}_{t+h} = U_{t+h+1} - U_{t+h}$.

The resulting post-price-movement, pre-rebalancing portfolio weights for the next imagined step are

$$w_{t+h+1}^{-} = \frac{w_{t+h} \odot \tilde{x}_{t+h+1}}{w_{t+h}^{\top} \tilde{x}_{t+h+1}}.$$

The imagined $H$-step return for one trajectory is therefore

$$J_t(\theta) = \sum_{h=0}^{H-1} \gamma^h \hat{r}_{t+h} + \gamma^H \mathcal{V}_\psi(\tilde{s}_{t+H}), \tag{1}$$

where $\mathcal{V}_\psi$ is the frozen critic used for the terminal bootstrap. For SAC, TD3, and DDPG, which learn action-value critics, the terminal value is $\mathcal{V}_\psi(\tilde{s}) = Q_\psi(\tilde{s}, \pi_\theta(\tilde{s}))$, taking the minimum over the twin critics for SAC and TD3.

**Risk-aware objective.** Because the imagined trajectory is stochastic under the forecast-error model, the inference-time objective is defined over its induced distribution. Let $\mu_t(\theta) = \mathbb{E}[J_t(\theta)]$ and define the downside variance $\mathcal{D}_t(\theta) = \mathbb{E}\left[\min\left(J_t(\theta) - \mu_t(\theta),\, 0\right)^2\right]$. We optimize

$$\mathcal{J}_t(\theta) = \mu_t(\theta) - \lambda \sqrt{\mathcal{D}_t(\theta) + \epsilon}. \tag{2}$$

In practice, we approximate these expectations using $K$ Monte Carlo forecast particles. For particle $k$, let $J_t^{(k)} = J_t(\theta; \tilde{s}^{(k)})$. We compute $\bar{J}_t = \frac{1}{K}\sum_{k=1}^{K} J_t^{(k)}$, and $\widehat{\mathcal{D}}_t = \frac{1}{K}\sum_{k=1}^{K} \min\left(J_t^{(k)} - \bar{J}_t,\, 0\right)^2$. The resulting Monte Carlo objective is $\widehat{\mathcal{J}}_t = \bar{J}_t - \lambda\sqrt{\widehat{\mathcal{D}}_t + \epsilon}$. In practice, the actor update uses Adam with gradient clipping, and both the terminal value and the particle mean $\bar{J}_t$ inside the downside term are treated as constants with respect to $\theta$. Consequently, the actor gradient flows only through the closed-form imagined PnL; the terminal bootstrap enters by shifting each particle's return, which determines which particles fall in the downside term and how strongly they are weighted.

## 4 Experimentation and Results

We apply `FinPILOT` to pre-trained policies obtained using PPO (Schulman et al., 2017), SAC (Haarnoja et al., 2018), A2C (Mnih et al., 2016), TD3 (Fujimoto et al., 2018), and DDPG (Lillicrap et al., 2019), five RL algorithms commonly used in RLFT benchmarks. For each trained checkpoint, the frozen baseline (Base) and `FinPILOT` start from identical parameters and are evaluated over the same test period, making every comparison paired. Base keeps its parameters fixed throughout the test period, whereas `FinPILOT` adapts the actor before each portfolio decision and carries the adapted parameters forward to the next decision. The adaptation uses a separately trained forecaster that is never observed by the policy during RL training.

Our experiments address three questions. **(Q1)** Does forecast-driven inference-time adaptation consistently improve the corresponding frozen Base policy across algorithms, test years, and two asset classes? **(Q2)** Given the same forecast, how does adapting a pre-trained policy compare with directly optimizing portfolio allocations? **(Q3)** How do the gains from inference-time adaptation change as forecast quality improves?

### 4.1 Experimental Setup

**Markets, protocol, and agents.** We use the TradeMaster portfolio-management environment (Sun et al., 2023) with two universes: 28 Dow Jones 30 constituents (DJ30) and 22 currency pairs quoted against the U.S. dollar (FX). Both use the same rolling protocol (Figure 2): for test year $Y \in$

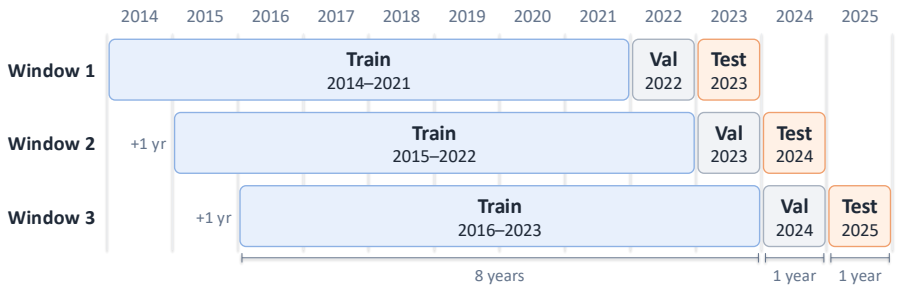



Figure 2: **Rolling evaluation protocol.** For each held-out test year $Y \in \{2023, 2024, 2025\}$, the preceding eight years are used for training and year $Y - 1$ for validation. The complete train/validation/test window advances by one year at each evaluation step. RL policies and forecasters are refit using only the corresponding training period.

$\{2023, 2024, 2025\}$, all models are trained on years $Y-9$ through $Y-2$, selected on $Y-1$, and evaluated once on $Y$. All five agents are pre-trained separately for each window, with batch size, hidden size, and learning rate selected on the validation year (Appendix C.1); no agent observes a forecast during training.

**Forecasters and `FinPILOT` configuration.** To reflect that `FinPILOT` is forecaster-agnostic, we use a different forecaster $F$ per market: a multi-horizon XGBoost model of per-step price differences for DJ30, and an equal-weight ensemble of a Theta model and a ridge-regularized VAR(1) for FX (Theta + VAR(1)-ridge; $R^2$ in Table 6). Only the risk weight $\lambda \in \{0.05, 0.10, 0.15\}$ is selected per market, algorithm, and window, by validation Sortino ratio; all other inference-time hyperparameters are shared across settings. The horizon $H = 50$ (also used by the multi-period baseline) was fixed during preliminary development on DJ30 data through 2021, before every validation and test year. Full settings are in Appendix C.2, and a post-hoc sensitivity analysis over step size and adaptation epochs, including settings that outperform ours on some metrics, is in Appendix D.4.

**Baselines.** Besides each agent's frozen policy (*Base*), we compare against two families of non-RL portfolios (Table 3). *Passive*: buy-and-hold and daily-rebalanced $1/N$. *Forecast-driven optimization*: single-period ($H$=1) and multi-period ($H$=50) convex trading optimizers (Boyd et al., 2017) that receive exactly the same forecasts as `FinPILOT`.

**Evaluation metrics.** Because the training reward captures only profit, we evaluate trained policies along five financial metrics standard in the RLFT literature (Sun et al., 2023; Wang et al., 2021; Ye et al., 2020; Liu et al., 2022a). We report Total Return (TR), Sharpe ratio (SR), Sortino ratio (SoR), Calmar ratio (CR), and Maximum Drawdown (MDD), following standard RL-for-trading evaluation practice. Formal definitions are provided in Appendix B.1. Unless otherwise stated, aggregate results average the three per-window seed means; complete per-window results with seed-level variability are reported in Tables 7 and 8 (Appendix D.1).

### 4.2 Improvement over Frozen RL Policies (Q1)

We first ask whether inference-time adaptation improves the static deployment of the same pre-trained RL policy. Table 1 reports performance averaged over the three rolling test windows. On DJ30, `FinPILOT` increases the three-window mean Total Return, Sharpe, Sortino, and Calmar ratios for all five pre-trained policies. The improvement is not restricted to stochastic-policy methods: TD3 and DDPG exhibit some of the largest return gains. These improvements are accompanied by higher mean Maximum Drawdown on DJ30, however, indicating that the larger realized returns do not uniformly translate into lower peak-to-trough risk. On FX, `FinPILOT` again improves the three-window mean Return, Sharpe, Sortino, and Calmar ratios for all five policies. In contrast to DJ30, mean Maximum Drawdown also decreases for all five policies, showing that the effect of inference-time adaptation on drawdown is market dependent.

Because averaging across windows can obscure whether a gain is driven by a single favorable period, Figure 3 reports per-window changes in Total Return and Sharpe ratio. Across the two markets,

Table 1: Performance averaged over the three rolling test windows (2023–2025). Each entry is the mean of the per-window seed means ($n = 5$ seeds per window). Values after $\pm$ are seed standard deviations pooled over the three windows (square root of the mean per-window seed variance), reflecting variability across seeds within a window, not across test years. Base is the frozen pre-trained policy; FinPILOT adapts its actor at inference time. Per-window results are in Figure 3 and Tables 7 and 8. Better mean per algorithm and metric in **bold**.

| | | DJ30 | | | | | FX | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Algo | Method | Ret. (%)↑ | Sharpe↑ | Sortino↑ | Calmar↑ | MDD (%)↓ | Ret. (%)↑ | Sharpe↑ | Sortino↑ | Calmar↑ | MDD (%)↓ |
| PPO | Base | 13.42 ±0.29 | 1.10 ±0.02 | 1.64 ±0.03 | 1.48 ±0.03 | **9.86** ±0.10 | 1.08 ±0.18 | 0.15 ±0.03 | 0.40 ±0.05 | 0.73 ±0.06 | 4.87 ±0.14 |
| | FinPILOT | **23.62** ±1.46 | **1.32** ±0.06 | **1.92** ±0.09 | **1.84** ±0.08 | 12.44 ±0.24 | **1.29** ±0.73 | **0.28** ±0.15 | **0.77** ±0.26 | **0.94** ±0.51 | **4.06** ±0.44 |
| SAC | Base | 13.77 ±0.05 | 1.13 ±0.00 | 1.68 ±0.00 | 1.51 ±0.00 | **9.79** ±0.01 | 1.20 ±0.06 | 0.16 ±0.01 | 0.42 ±0.01 | 0.75 ±0.01 | 4.80 ±0.03 |
| | FinPILOT | **15.95** ±0.04 | **1.22** ±0.00 | **1.89** ±0.00 | **1.73** ±0.00 | 9.89 ±0.01 | **1.29** ±0.04 | **0.20** ±0.01 | **0.58** ±0.02 | **0.85** ±0.01 | **4.38** ±0.01 |
| A2C | Base | 13.73 ±0.10 | 1.12 ±0.01 | 1.67 ±0.01 | 1.50 ±0.01 | **9.81** ±0.04 | 1.20 ±0.02 | 0.16 ±0.00 | 0.42 ±0.01 | 0.75 ±0.01 | 4.81 ±0.01 |
| | FinPILOT | **27.17** ±0.69 | **1.45** ±0.03 | **2.23** ±0.04 | **2.46** ±0.10 | 10.59 ±0.14 | **1.44** ±0.15 | **0.33** ±0.03 | **0.85** ±0.05 | **0.79** ±0.03 | **3.92** ±0.11 |
| TD3 | Base | 12.93 ±4.81 | 0.99 ±0.38 | 1.43 ±0.46 | 1.25 ±0.62 | **11.45** ±1.73 | 0.54 ±1.11 | 0.09 ±0.19 | 0.30 ±0.26 | 0.72 ±0.44 | 5.19 ±0.83 |
| | FinPILOT | **36.30** ±9.41 | **1.55** ±0.34 | **2.49** ±0.60 | **2.55** ±0.71 | 14.21 ±3.84 | **3.42** ±1.44 | **0.53** ±0.25 | **1.06** ±0.37 | **1.21** ±0.20 | **4.47** ±1.09 |
| DDPG | Base | 10.80 ±4.30 | 0.85 ±0.31 | 1.23 ±0.45 | 1.08 ±0.54 | **11.40** ±1.70 | −0.21 ±1.66 | −0.11 ±0.33 | −0.05 ±0.56 | 0.28 ±0.59 | 5.19 ±1.05 |
| | FinPILOT | **36.19** ±7.78 | **1.55** ±0.32 | **2.64** ±0.68 | **2.73** ±0.75 | 13.04 ±2.18 | **1.48** ±1.83 | **0.19** ±0.39 | **0.51** ±0.57 | **0.33** ±0.33 | **5.00** ±1.04 |

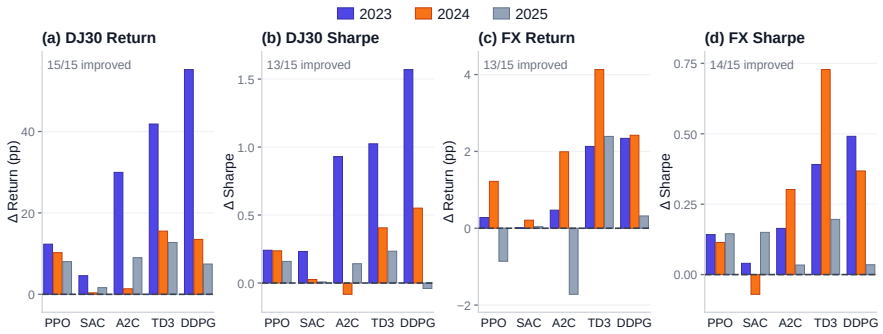



Figure 3: **Per-window improvement from FinPILOT.** Bars show the change relative to the corresponding frozen Base policy for each RL policy and held-out test year. Positive values indicate improvement from inference-time adaptation. Colors denote test years. Across the 15 algorithm–window pairs in each market, Total Return improves in 15/15 DJ30 and 13/15 FX cases, while Sharpe ratio improves in 13/15 and 14/15 cases, respectively.

FinPILOT improves Total Return in **28 of 30** and Sharpe ratio in **27 of 30** policy–market–window comparisons, showing that the aggregate gains are broadly distributed across evaluation periods. Complete per-window results with seed-level variability are reported in Appendix D.1.

### 4.3 The same forecast, consumed differently (Q2)

Given the same forecast, one can bypass the pre-trained RL policy and optimize allocations directly. We compare FinPILOT with single-period optimization (SPO) and multi-period optimization (MPO), which optimize the current allocation and an $H$-step allocation sequence, respectively. FinPILOT instead optimizes the parameters of the pre-trained actor (Table 2).

Table 2: Comparison of inference-time forecast integration mechanisms.

| Method | Pre-trained RL | Forecast | Optimization target | Horizon |
|---|---|---|---|---|
| Base RL | ✓ | ✗ | None | – |
| SPO | ✗ | ✓ | Current allocation | 1 |
| MPO | ✗ | ✓ | Allocation sequence | $H$ |
| FinPILOT | ✓ | ✓ | Policy parameters | $H$ |

Under this evaluation protocol, directly optimizing allocations from the forecast does not reproduce the aggregate performance obtained by adapting the pre-trained policies (Table 3). On DJ30, every FinPILOT-adapted policy achieves higher mean Total Return, Sharpe ratio, and Sortino ratio than both SPO and MPO. On FX, SPO and MPO select essentially cash-only portfolios and therefore obtain near-zero mean returns, whereas the adapted RL policies remain exposed to the market and

Table 3: Comparison with non-RL baselines, averaged over the three rolling test windows (2023–2025). FinPILOT entries are the mean of the per-window seed means ($n = 5$ seeds per window); all baselines are deterministic. SPO and MPO receive the same forecasts as FinPILOT, whereas the passive allocators use no forecast. Best value per column within each market in **bold** (values tied at two decimals are all bolded). †On FX, the Boyd optimizers remain essentially in cash in every window, so their near-zero returns and drawdowns reflect inactivity rather than skill; they are excluded when marking the best MDD, and their Sharpe and Sortino ratios, computed from near-zero returns, are numerically unstable. Per-window results are in Tables 9 and 10.

| | DJ30 | | | | FX | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Return (%) ↑ | Sharpe ↑ | Sortino ↑ | MDD (%) ↓ | Return (%) ↑ | Sharpe ↑ | Sortino ↑ | MDD (%) ↓ |
| Buy-and-hold | 14.70 | 1.18 | 1.77 | 10.02 | 1.31 | 0.17 | 0.43 | 5.01 |
| $1/N$ (daily rebalanced) | 14.25 | 1.13 | 1.68 | 10.12 | 1.26 | 0.16 | 0.42 | 5.02 |
| *Direct optimization* | | | | | | | | |
| Boyd SPO ($H$=1) | 8.95 | 0.52 | 0.63 | 17.62 | 0.00† | −0.71 | −0.80 | <0.01† |
| Boyd MPO ($H$=50) | 6.25 | 0.37 | 0.49 | 19.14 | 0.00† | −2.98 | −3.51 | <0.01† |
| *Policy optimization* | | | | | | | | |
| FinPILOT-PPO | 23.62 | 1.32 | 1.92 | 12.44 | 1.29 | 0.28 | 0.77 | 4.06 |
| FinPILOT-SAC | 15.95 | 1.22 | 1.89 | **9.89** | 1.29 | 0.20 | 0.58 | 4.38 |
| FinPILOT-A2C | 27.17 | 1.45 | 2.23 | 10.59 | 1.44 | 0.33 | 0.85 | **3.92** |
| FinPILOT-TD3 | **36.30** | **1.55** | 2.49 | 14.21 | **3.42** | **0.53** | **1.06** | 4.47 |
| FinPILOT-DDPG | 36.19 | **1.55** | **2.64** | 13.04 | 1.48 | 0.19 | 0.51 | 5.00 |

achieve positive three-window mean returns. These results should not be interpreted as a general claim that policy optimization dominates direct portfolio optimization. Rather, they show that access to the forecast alone is not sufficient to reproduce the gains of FinPILOT under our evaluation protocol: the mechanism by which the forecast is incorporated into the decision rule matters.

### 4.4 How much forecast skill is needed (Q3)

A natural question is how much further the framework could go with a stronger forecaster. To answer this, we construct controlled forecasts at calibrated $R^2$ levels.

**Controlled forecast-skill condition.** To vary forecast quality while preserving the horizon-wise error pattern of the forecasters used in our main experiments, we rescale their errors. For origin $t$ and asset $i$, let $\hat{d}_{t,i,h}$ denote the original forecast and $d_{t,i,h}$ the realized one-day price change at step $t+h$, $h = 1, \ldots, H$. We set $\tilde{d}_{t,i,h} = d_{t,i,h} + c_{t,i}(\hat{d}_{t,i,h} - d_{t,i,h})$ with $c_{t,i} = \sqrt{(1 - R^2_\star)/(1 - R^2_{t,i})}$, where $R^2_{t,i}$ is the original forecast's $R^2$ over the $H$-step horizon relative to a trailing-mean baseline (Appendix C.3). This scales the squared error by $c^2_{t,i}$ without changing the baseline, so each controlled forecast attains exactly $R^2_\star$ at every origin and asset. Targets below the original skill ($c_{t,i} > 1$) amplify the original errors, while targets above it ($c_{t,i} < 1$) move the forecast toward the realized path, with $c_{t,i} = 1$ and $c_{t,i} = 0$ recovering the original forecast and the realized path, respectively. Because it uses realized future prices, this construction is an oracle diagnostic for mapping forecast skill to planning performance, not a deployable forecaster.

**Scaling with forecast skill.** Figure 4 shows that stronger predictive signals produce both larger and more consistent gains from inference-time adaptation. In the weakest controlled-signal regime, adaptation is not uniformly beneficial across policies and market windows. As $R^2_\star$ increases, however, the mean Sharpe improvement rises in both markets and the fraction of improved settings increases correspondingly. By $R^2_\star = 0.05$, all 15 policy–window comparisons in each market improve over Base in Sharpe, and this remains true for the stronger controlled forecasts shown in Figure 4. We do not read $R^2_\star = 0.05$ as a practical threshold: such skill is hard to attain in financial prediction, the construction is oracle-based, and forecasts with similar $R^2$ can differ in cross-asset, temporal, and directional error structure.

**Original vs. controlled forecasts.** The original forecasts operate in the low-skill region of the controlled sweep. Their mean Sharpe improvements are close to those obtained with controlled

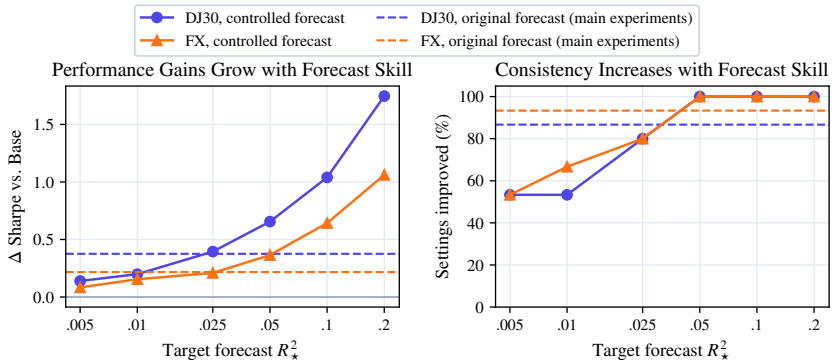



Figure 4: **Performance scales with controlled forecast skill.** Left: mean Sharpe improvement over the corresponding frozen Base policy as target forecast quality $R^2_\star$ increases. Right: percentage of the 15 policy–window settings per market in which FinPILOT improves Sharpe. Solid curves denote controlled forecasts, while dashed horizontal lines show the corresponding results obtained with the original forecasts used in the main experiments. Controlled forecasts use realized future prices by construction and are therefore diagnostic rather than deployable.

forecasts at $R^2_\star = 0.025$: $+0.38$ versus $+0.39$ on DJ30 and $+0.22$ versus $+0.21$ on FX. Yet with the original forecasts, FinPILOT improves Sharpe more consistently, in 13/15 and 14/15 policy–window comparisons on DJ30 and FX, respectively, compared with 12/15 for the controlled $R^2_\star = 0.025$ condition in each market. This is consistent with the controlled construction enforcing the same $R^2_\star$ at every origin and asset, whereas the original forecaster's skill varies, and suggests that aggregate $R^2$ alone does not fully characterize decision value.

Taken together, these results answer Q3: FinPILOT already extracts useful decision signals from the low-skill forecasters used in the main experiments, while stronger controlled predictive signals produce larger and more consistent gains. Because the policy and adaptation procedure are held fixed throughout this sweep, the result supports the interpretation that the improvements arise from predictive information in the forecast rather than simply from performing additional gradient updates at inference time.

## 5 Conclusion

We introduced FinPILOT, an inference-time policy adaptation framework that uses an independently trained financial forecaster to adapt a pre-trained RL trading policy before each decision. Across five pre-trained policies, two markets, and three rolling held-out test years, FinPILOT improves Total Return in 28 of 30 policy–market–window comparisons and Sharpe ratio in 27 of 30. Same-forecast direct portfolio optimizers do not reproduce these aggregate gains, while controlled forecast-skill experiments show that the magnitude and consistency of the benefit increase with predictive signal quality. These results support inference-time policy adaptation as a modular interface between separately trained forecasters and pre-trained RL trading policies.

## Limitations and Future Work

Our study is limited to historical backtests on DJ30 and FX under the price-taker assumption, which neglects the market impact of large or aggressive trades. The formulation also assumes long-only, unlevered, daily portfolios. Our evaluation covers only three recent test years (2023–2025), so broader validation across assets, market regimes, frequencies, and portfolio constraints is needed. The benefit of adaptation depends on its magnitude: we fix the step size and epoch count across all settings, but larger updates help when the forecast is informative and hurt when it is weak (e.g., DJ30 2024; Appendix D.4), so no single setting is best in every market and year. The controlled forecast-skill experiments use realized future prices and are diagnostics, not deployable results. Future work could extend FinPILOT to market-impact-aware environments, richer portfolio constraints and risk objectives, update magnitudes that adapt to estimated forecast skill, and probabilistic forecasters.

## AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, or interpreting results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. AI-assisted code was reviewed, tested, and verified for correctness by the authors, and AI-suggested edits to the manuscript were limited to language and clarity and were checked by the authors to ensure they did not alter any technical content or claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

This work studies inference-time adaptation of reinforcement learning agents for portfolio management on historical market data. It involves no human subjects or personal data; all experiments use publicly available daily market prices (Yahoo Finance and the Federal Reserve H.10 release) for non-commercial research. As with other automated-trading research, wide deployment of similar strategies could contribute to herding and market volatility, and ML-based trading tools tend to benefit well-resourced institutional actors more than retail investors; our method offers no specific advantage for manipulative strategies over existing approaches. All results are historical backtests: backtested gains do not reliably translate to live trading because of market impact, slippage, latency, and regime shifts that our environment does not model, and we do not deploy live trading systems. The released code is intended for research and reproducibility, and reported gains should be read as benchmark-relative rather than as forecasts of real-world profitability.

## REPRODUCIBILITY STATEMENT

We release anonymized code and data as supplementary material: the daily price files for both datasets, the full implementation of the inference-time procedure (Algorithm 1, expanded in Algorithm 2) for all five RL algorithms, and scripts that rebuild the train/validation/test windows, pretrain the RL agents (including the per-window hyperparameter search), train the forecasters, run the non-RL baselines, and generate the controlled-skill forecasts. Full experimental settings are given in Appendices B and C.1–C.4. All experiments follow the DJ30 and FX asset universes of TradeMaster (Sun et al., 2023), extended through 2025. Price data are provided for non-commercial research use, subject to the data providers' terms.

## REFERENCES


Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making?, 2023. URL https://arxiv.org/abs/2211.15657.

Stephen P. Boyd, Enzo Busseti, Steven Diamond, Ronald N. Kahn, Kwangmoo Koh, Peter Nystrup, and Jan Speth. Multi-period trading via convex optimization. *Found. Trends Optim.*, 3(1):1–76, 2017. doi: 10.1561/2400000023. URL https://doi.org/10.1561/2400000023.

John Y. Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531, 07 2008. ISSN 0893-9454. doi: 10.1093/rfs/hhm055. URL https://doi.org/10.1093/rfs/hhm055.

Rohan Deb, Stephen J. Wright, and Arindam Banerjee. Inference time policy optimization for offline rl with differentiable world models, 2026. URL https://arxiv.org/abs/2603.22430.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training, 2019. URL https://arxiv.org/abs/1810.09936.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Carlos E. García, David M. Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(89)90002-2. URL https://www.sciencedirect.com/science/article/pii/0005109889900022.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 02 2020. ISSN 0893-9454. doi: 10.1093/rfs/hhaa009. URL https://doi.org/10.1093/rfs/hhaa009.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5faf461eff3099671ad63c6f3f094f7f-Paper.pdf.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1273–1286. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/099fe6b0b444c23836c4a5d07346082b-Paper.pdf.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem, 2017. URL https://arxiv.org/abs/1706.10059.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhaoran Wang, and Jian Guo. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1835–1849. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0bf54b80686d2c4dc0808c2e98d430f7-Paper-Datasets_and_Benchmarks.pdf.

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, 2022b. URL https://arxiv.org/abs/2011.09607.

D.Q. Mayne, J.B. Rawlings, C.V. Rao, and P.O.M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000. ISSN 0005-1098. doi: https://doi.org/10.1016/S0005-1098(99)00214-9. URL https://www.sciencedirect.com/science/article/pii/S0005109899002149.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/mniha16.html.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, Dec 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-03051-4. URL https://doi.org/10.1038/s41586-020-03051-4.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

William F. Sharpe. The Sharpe ratio. *Journal of Portfolio Management*, 21(1):49–58, 1994.

Shuo Sun, Wanqi Xue, Rundong Wang, Xu He, Junlei Zhu, Jian Li, and Bo An. Deepscalper: A risk-aware reinforcement learning framework to capture fleeting intraday trading opportunities. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 1858–1867, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557283. URL https://doi.org/10.1145/3511808.3557283.

Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. Trademaster: A holistic quantitative trading platform empowered by reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 59047–59061. Curran Associates, Inc., 2023.

URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b8f6f7f2ba4137124ac976286eacb611-Paper-Datasets_and_Benchmarks.pdf.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. Deeptrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):643–650, May 2021. doi: 10.1609/aaai.v35i1.16144. URL https://ojs.aaai.org/index.php/AAAI/article/view/16144.

Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, Ju Xiao, and Bo Li. Reinforcement-learning based portfolio management with augmented asset movement prediction states. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1112–1119, Apr. 2020. doi: 10.1609/aaai.v34i01.5462. URL https://ojs.aaai.org/index.php/AAAI/article/view/5462.

Jaemin Yoo, Yejun Soun, Yong-chan Park, and U Kang. Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2037–2045, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467297. URL https://doi.org/10.1145/3447548.3467297.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Wentao Zhang, Yilei Zhao, Shuo Sun, Jie Ying, Yonggang Xie, Zitao Song, Xinrun Wang, and Bo An. Reinforcement learning with maskable stock representation for portfolio management in customizable stock pools. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 187–198, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645615. URL https://doi.org/10.1145/3589334.3645615.

# A Related Work

**Reinforcement Learning for Financial Trading.** The application of RL to portfolio management has grown substantially, with early tabular methods giving way to deep RL approaches. Jiang et al. (2017) made the first deep RL attempt on portfolio management, and subsequent work introduced risk-aware objectives and augmented state representations incorporating price prediction signals (Wang et al., 2021; Sun et al., 2022; Ye et al., 2020). Sun et al. (2023) and Liu et al. (2022b) consolidated these into reproducible benchmarking platforms, which we build on directly. A consistent limitation across all of these methods is that the trained policy is deployed statically; there is no mechanism for exploiting a price forecaster at inference time within a principled planning framework. Our work addresses this gap.

**Model Predictive Control and World Models.** Model-based RL methods such as MBPO and MOPO use learned world models during training to generate imagined rollouts for policy improvement (Janner et al., 2019; Yu et al., 2020). A separate line of work performs inference-time planning: Trajectory Transformer (Janner et al., 2021), Decision Diffuser (Ajay et al., 2023), and TD-MPC2 (Hansen et al., 2024) sample candidate plans at test time but do not backpropagate through the rollout to adapt policy parameters. The most directly relevant work is Deb et al. (2026), which introduced inference-time gradient-based policy adaptation through a learned differentiable world model on offline RL benchmarks. We extend this paradigm to financial trading. As we argue in Section 1, the structure of portfolio management simplifies the world modeling requirements substantially, reducing the problem to price forecasting alone.

**Model Predictive Control (MPC) in Finance.** Boyd et al. (2017) established convex optimization-based multi-period trading as the classical MPC framework in quantitative finance, optimizing action sequences over predicted future returns subject to transaction cost penalties. Our work connects this tradition to learned RL policies, replacing the convex planner with gradient-based adaptation of a pre-trained actor.

**Price Forecasting as a Trading Signal.** Practical price movement prediction for equities is a well-studied but fundamentally difficult problem. Even state-of-the-art machine learning methods achieve modest out-of-sample $R^2$ values for short-horizon equity prediction (Gu et al., 2020). Our XGBoost forecaster for DJ30 predicts daily price movements rather than returns directly. On daily close-price moves over the $H = 50$ horizon, averaged across test years, it achieves a pooled out-of-sample $R^2$ of 0.016, and the Theta + VAR(1)-ridge ensemble for FX achieves 0.029 (Table 6). While modest in absolute terms, out-of-sample $R^2$ of this magnitude can still carry meaningful economic value (Campbell & Thompson, 2008).

# B Temporal Feature Construction & Metrics

For each asset $i$ at trading day $t$, we construct an 11-dimensional feature vector following the input representation used in TradeMaster (Sun et al., 2023), adapted from prior work on stock movement prediction (Feng et al., 2019; Yoo et al., 2021). The features summarize short-term price action and longer-horizon momentum entirely from the daily OHLC and adjusted-close series, requiring no auxiliary data (Table 4).

The four intra-day features encode the open, high, and low prices relative to the same day's close, together with the adjusted close, which corrects for corporate actions such as dividends and splits:

$$z_{\text{open}} = \frac{\text{open}_t}{\text{close}_t} - 1, \quad z_{\text{high}} = \frac{\text{high}_t}{\text{close}_t} - 1, \quad z_{\text{low}} = \frac{\text{low}_t}{\text{close}_t} - 1, \quad z_{\text{adj}} = \frac{\text{adj_close}_t}{\text{close}_t} - 1. \tag{3}$$

A single one-day return captures day-over-day price change:

$$z_{\text{close}} = \frac{\text{close}_t}{\text{close}_{t-1}} - 1. \tag{4}$$

Six moving-average ratios capture momentum at horizons $k \in \{5, 10, 15, 20, 25, 30\}$ trading days, expressing the average close over the last $k$ days relative to today's close:

$$z_{d_k} = \frac{\frac{1}{k}\sum_{j=0}^{k-1} \text{close}_{t-j}}{\text{close}_t} - 1. \tag{5}$$

All 11 features are dimensionless ratios relative to the same day's or the previous day's close, so they are comparable across assets without further scaling. We apply no additional normalization.

Table 4: Summary of the 11 temporal features computed per asset per day.

| Feature | Definition |
|---|---|
| $z_{\text{open}}$ | $\text{open}_t/\text{close}_t - 1$ |
| $z_{\text{high}}$ | $\text{high}_t/\text{close}_t - 1$ |
| $z_{\text{low}}$ | $\text{low}_t/\text{close}_t - 1$ |
| $z_{\text{adj}}$ | $\text{adj_close}_t/\text{close}_t - 1$ |
| $z_{\text{close}}$ | $\text{close}_t/\text{close}_{t-1} - 1$ |
| $z_{d_k}$, $k \in \{5, 10, 15, 20, 25, 30\}$ | $\left(\frac{1}{k}\sum_{j=0}^{k-1} \text{close}_{t-j}\right)/\text{close}_t - 1$ |

### B.1 Evaluation Metric Definitions

Let $U_t$ denote portfolio value at time $t$, $T$ the test horizon length, and $\rho_t = U_t/U_{t-1} - 1$ the daily portfolio return. We annualize with 252 trading days per year and assume a zero risk-free rate, consistent with the TradeMaster benchmark.

**Total Return (TR).**

$$\text{TR} = \frac{U_T - U_0}{U_0} \tag{6}$$

**Sharpe Ratio (SR) (Sharpe, 1994).** Annualized ratio of mean excess return to return volatility:

$$\text{SR} = \sqrt{252} \cdot \frac{\bar{\rho}}{\sigma_\rho}, \quad \bar{\rho} = \tfrac{1}{T}\sum_{t=1}^{T} \rho_t, \quad \sigma_\rho = \sqrt{\tfrac{1}{T-1}\sum_{t=1}^{T}(\rho_t - \bar{\rho})^2} \tag{7}$$

**Sortino Ratio (SoR).** Annualized ratio of mean return to downside deviation:

$$\text{SoR} = \sqrt{252} \cdot \frac{\bar{\rho}}{\sigma_d}, \quad \sigma_d = \sqrt{\tfrac{1}{T}\sum_{t=1}^{T} \min(\rho_t, 0)^2} \tag{8}$$

**Maximum Drawdown (MDD).** Worst peak-to-trough decline over the test period:

$$\text{MDD} = \max_{t \in [0,T]} \left( \frac{\max_{s \le t} U_s - U_t}{\max_{s \le t} U_s} \right) \tag{9}$$

**Calmar Ratio (CR).** Annualized return divided by maximum drawdown:

$$\text{CR} = \frac{(1 + \text{TR})^{252/T} - 1}{\text{MDD}} \tag{10}$$

## C Experimental Setup

### C.1 Training Details and Hyperparameter Selection

**Baseline RL training.** The TradeMaster paper does not release the specific hyperparameters used for their reported baselines, so we re-tuned each algorithm on our setup. Following the search space described in the TradeMaster paper, we grid-searched batch size in $\{256, 512, 1024\}$, hidden size in $\{64, 128, 256\}$, and actor/critic learning rate in $\{3\text{e}{-4}, 5\text{e}{-4}, 7\text{e}{-4}, 9\text{e}{-4}\}$ for each of PPO, SAC, A2C, TD3, and DDPG, with Adam as the optimizer in all cases. Selection was based on the best total return on the validation split. The selected configuration was then retrained across 5 random seeds to produce the reported baseline numbers.

**Algorithm 2** FinPILOT: Inference-Time Policy Adaptation for Portfolio Management

**Require:** Pre-trained actor $\pi_\theta$, frozen critic $\mathcal{V}_\psi$, forecaster $F$, horizon $H$, particles $K$, epochs $E$, step size $\alpha$, discount $\gamma$, risk weight $\lambda$, noise scale $\sigma$

**for** $t = 1, \dots, T$ **do**
  Observe real state $s_t$, pre-rebalancing portfolio $w_t^-$, and wealth $U_t$
  Predict the $H$-step state trajectory $\{\hat{s}_{t+h}\}_{h=1}^{H} \leftarrow F(s_{\le t})$ ▷ forecaster frozen
  **for** $k = 1, \dots, K$ **do**
    **for** $h = 1, \dots, H$ **do**
      Sample forecast error $\varepsilon_{t,h}^{(k)} \sim \mathcal{N}(0, \sigma^2 \Sigma_h^{\mathrm{val}})$
      Construct imagined state $\tilde{s}_{t+h}^{(k)} \leftarrow \mathrm{Perturb}(\hat{s}_{t+h}, \varepsilon_{t,h}^{(k)})$
  ▷ Particles are fixed for all $E$ adaptation epochs
  **for** $e = 1, \dots, E$ **do**
    **for** $k = 1, \dots, K$ **do**
      $J_t^{(k)} \leftarrow J_t\big(\theta; \tilde{s}_{t:t+H}^{(k)}, w_t^-, U_t\big)$ ▷ equation 1
    $\bar{J}_t \leftarrow \frac{1}{K} \sum_{k=1}^{K} J_t^{(k)}$
    $\widehat{\mathcal{D}}_t \leftarrow \frac{1}{K} \sum_{k=1}^{K} \min\left(J_t^{(k)} - \bar{J}_t,\, 0\right)^2$
    $\widehat{\mathcal{J}}_t \leftarrow \bar{J}_t - \lambda \sqrt{\widehat{\mathcal{D}}_t + \epsilon}$
    $\theta \leftarrow \theta + \alpha \nabla_\theta \widehat{\mathcal{J}}_t$ ▷ actor parameters only
  Execute $w_t = \pi_\theta(s_t)$ in the real environment

### C.2 FinPILOT CONFIGURATION

Table 5 lists all inference-time settings. Only the risk weight $\lambda$ is selected per setting; every other value is shared across algorithms, markets, and windows.

**Shared settings.** The planning horizon was chosen from $H \in \{1, 15, 50\}$ during preliminary development on DJ30 data through 2021, which precedes every validation and test year in our protocol, and fixed at $H = 50$ thereafter. The step size ($\alpha = 0.01$), number of adaptation epochs ($E = 2$), and number of forecast particles ($K = 512$) are single values shared across all settings, with no per-setting tuning. Imagined rollouts start from the agent's current portfolio value (initial capital $10^5$, compounding over the test period), the same reward scale on which the critic was trained, so imagined rewards and the terminal bootstrap share the same currency units. The actor is updated with Adam, with gradients clipped to unit norm. The discount $\gamma = 0.99$ matches pretraining; with $H = 50$, the terminal bootstrap receives weight $\gamma^H \approx 0.61$. The forecast-noise covariance $\Sigma_h^{\mathrm{val}}$ is estimated from forecast errors on the validation split of each window and used without rescaling ($\sigma = 1$), so the dispersion of imagined trajectories reflects the forecaster's observed error rather than a tuned noise level.

**Risk-weight selection.** For each market, algorithm, and window, we run FinPILOT on the validation year with each $\lambda \in \{0.05, 0.10, 0.15\}$ and select the value with the highest mean validation Sortino ratio across seeds. The selected $\lambda$ is held fixed for the corresponding test year. DJ30 and FX are tuned independently under the same protocol. We select only $\lambda$ per setting because it encodes a risk preference, namely the trade-off between imagined return and downside dispersion, whose appropriate level can differ across policies with different base risk profiles and across markets. We use the Sortino ratio rather than total return for this choice because $\lambda$ penalizes downside dispersion, so a downside-risk-adjusted criterion matches its role. A post-hoc sensitivity analysis over the step size and number of adaptation epochs is given in Appendix D.4.

### C.3 FORECASTERS AND FORECAST SKILL

**Definition of $R_{t,i}^2$.** The controlled forecasts in Section 4.4 rescale errors using the original forecast's skill $R_{t,i}^2$ at each origin $t$ and asset $i$. Let $d_{s,i}$ denote the realized one-day price change of asset $i$ on day $s$, so that $d_{t,i,h} = d_{t+h,i}$ in the notation of Section 4.4. The baseline is the trailing

Table 5: FinPILOT inference-time hyperparameters. $\lambda$ is chosen on each window's validation year; $H$ was fixed during development on pre-2022 data; all other FinPILOT settings are shared and untuned.

| Symbol | Description | Value | How set |
|---|---|---|---|
| $H$ | Planning horizon (days) | 50 | Prelim. dev. (DJ30, $\leq$2021) |
| $\alpha$ | Actor step size | 0.01 | Shared |
| $E$ | Adaptation epochs per decision | 2 | Shared |
| $K$ | Forecast particles | 512 | Shared |
| $\gamma$ | Discount factor | 0.99 | Matches pretraining |
| $\sigma$ | Noise scale on $\Sigma_h^{\text{val}}$ | 1 | Unscaled validation cov. |
| $\epsilon$ | Stabilizer in downside penalty | $10^{-8}$ | Numerical constant |
| $c_{\text{tx}}$ | Transaction cost | 10 bps | Environment |
| $U_0$ | Initial capital | $10^5$ | Environment |
| – | Actor optimizer | Adam | – |
| – | Gradient-norm clip | 1 | – |
| $\lambda$ | Risk weight | $\{0.05, 0.10, 0.15\}$ | Val. Sortino, per setting |

mean of realized changes available at the origin,

$$\bar{d}_{t,i} = \frac{1}{L} \sum_{j=1}^{L} d_{t-j,i}, \tag{11}$$

with $L = 30$ trading days, held constant across the horizon. The skill is

$$R^2_{t,i} = 1 - \frac{\sum_{h=1}^{H} \left(\hat{d}_{t,i,h} - d_{t,i,h}\right)^2}{\sum_{h=1}^{H} \left(\bar{d}_{t,i} - d_{t,i,h}\right)^2}. \tag{12}$$

The construction requires $R^2_{t,i} < 1$, which holds at every origin in our data. Table 6 reports this skill for each forecaster and test year, pooled over forecast origins, horizons, and assets.

Table 6: Out-of-sample $R^2$ of the learned forecasters on each test year. $R^2$ is computed on daily close-price moves over the $H = 50$ planning horizon, relative to the trailing 30-day mean move, and pooled over forecast origins, horizons, and assets.

| | | Test year | | | |
|---|---|---|---|---|---|
| Market | Forecaster | 2023 | 2024 | 2025 | Mean |
| DJ30 | XGBoost | 0.027 | 0.005 | 0.014 | 0.016 |
| FX | Theta + VAR(1)-ridge | 0.039 | 0.029 | 0.019 | 0.029 |

### C.4 Non-RL baselines

All baselines run in the same environment as FinPILOT, with the same transaction cost ($c_{\text{tx}} = 10$ bps on $\ell_1$ turnover).

**Passive allocators.** *Buy-and-hold* invests $1/N$ in each of the $N$ risky assets on the first day of the test year and never trades again. *1/N* rebalances to equal weights across the $N$ risky assets every day. Neither holds cash or uses a forecast.

**Forecast-driven optimizers.** Following Boyd et al. (2017), adapted to our long-only, unlevered setting with the environment's linear transaction cost (no holding cost or market-impact term), the single-period (SPO, $H = 1$) and multi-period (MPO, $H = 50$, matching FinPILOT's planning

horizon) optimizers choose weights over cash and the $N$ assets by solving, at each decision step $t$,

$$\max_{w_1,\ldots,w_H} \sum_{h=1}^{H} \left( \hat{\mu}_{t,h}^\top w_h - \gamma_{\text{risk}}\, w_h^\top \hat{\Sigma}_t\, w_h - \gamma_{\text{trade}}\, c_{\text{tx}} \|w_h - w_{h-1}\|_1 \right) \quad \text{s.t. } w_h \geq 0,\ \mathbf{1}^\top w_h = 1, \tag{13}$$

with $w_0$ the current pre-rebalancing weights, where cash has zero expected return and zero risk, and execute only $w_1$ before replanning at the next step. The expected returns $\hat{\mu}_{t,h}$ are the predicted one-day close-price returns at horizon $h$, taken from exactly the same forecasts that FinPILOT receives. The planner ignores weight drift within the horizon; the environment applies the realized dynamics. The risk model $\hat{\Sigma}_t$ is the factor model of Boyd et al. (2017, Sec. 7.3) with 15 factors, estimated from the trailing 500 trading days of realized daily returns and re-estimated on the first trading day of each month, using only returns realized up to that date. Problems are solved with CVXPY using OSQP.

**Hyperparameter selection.** We search 87 distinct pairs $(\gamma_{\text{risk}}, \gamma_{\text{trade}})$, the union of the coarse grid $\{0.1, 1, 10, 100, 1000\} \times \{1, 2, 5, 10, 20\}$ and the finer grid $\{0.03, 0.056, 0.1, 0.178, 0.316, 0.562, 1, 1.78\} \times \{7, 8, 9, 12, 15, 20, 30, 50\}$. For each test year $Y$, we select the pair with the highest mean final portfolio value over the years preceding it (2022 through $Y-1$), each evaluated with a forecaster trained only on data preceding that year; no result from year $Y$ is used. Ties go to the smaller $\gamma_{\text{risk}}$, then the smaller $\gamma_{\text{trade}}$.

**Cash-only solutions on FX.** On FX, the forecast-implied expected returns are small relative to the trading penalty: every grid pair with $\gamma_{\text{trade}} \geq 7$ remains essentially in cash (final value within 0.01% of the initial value) in all three test years, and the pairs selected on the preceding years all have $\gamma_{\text{trade}} \geq 20$, reflecting losses from trading in 2022, the first selection year.

# D Additional Experimental Results

This appendix reports the per-window results behind the three-window averages in the main text (Appendices D.1 and D.2), the full results of the controlled forecast-skill experiments (Appendix D.3), and a post-hoc sensitivity analysis over the step size and number of adaptation epochs (Appendix D.4).

## D.1 Per-Window Comparison with Frozen Policies

Tables 7 and 8 report per-window results underlying the three-window averages in Table 1 and the per-window changes in Figure 3. On DJ30, FinPILOT improves total return over Base in all 15 algorithm–windows and Sharpe in 13; in the two exceptions (A2C in 2024 and DDPG in 2025), return still improves while Sharpe falls slightly. Maximum drawdown increases in 12 of the 15 DJ30 algorithm–windows, consistent with Section 4.2. On FX, total return improves in 13 of 15 algorithm–windows and Sharpe in 14. PPO and A2C in 2025 lose return but improve Sharpe and reduce drawdown, and SAC in 2024 gains return but loses Sharpe. In 2024, when all five base policies lost between 5.7% and 7.4%, FinPILOT reduced both the loss and the maximum drawdown for every algorithm (Table 8). Because each FinPILOT run starts from the Base policy with the same seed, comparisons are paired; overlap of the $\pm$ ranges does not by itself imply that a difference is not significant.

Table 7: Per-window performance on DJ30 (three rolling test windows, XGBoost forecaster). Values are mean ± one standard deviation across random seeds ($n = 5$ per window cell). Base denotes the frozen pre-trained policy; FinPILOT denotes the same agent under our method. For each window, algorithm, and metric, the better of the two means is shown in **bold** (higher is better for all metrics except MDD, where lower is better); values tied at two decimals are both bolded.

| Window | Algo | Method | Return (%) | Sharpe | Sortino | Calmar | MDD (%) |
|---|---|---|---|---|---|---|---|
| 2023 | PPO | Base | $13.84 \pm 0.39$ | $1.25 \pm 0.02$ | $2.07 \pm 0.04$ | $1.64 \pm 0.04$ | $\mathbf{8.27 \pm 0.05}$ |
| | | FinPILOT | $\mathbf{26.16 \pm 0.43}$ | $\mathbf{1.50 \pm 0.02}$ | $\mathbf{2.34 \pm 0.04}$ | $\mathbf{2.03 \pm 0.03}$ | $12.11 \pm 0.00$ |
| | SAC | Base | $14.25 \pm 0.01$ | $1.28 \pm 0.00$ | $2.12 \pm 0.00$ | $1.68 \pm 0.00$ | $\mathbf{8.30 \pm 0.00}$ |
| | | FinPILOT | $\mathbf{18.82 \pm 0.01}$ | $\mathbf{1.51 \pm 0.00}$ | $\mathbf{2.61 \pm 0.00}$ | $\mathbf{2.01 \pm 0.00}$ | $8.94 \pm 0.00$ |
| | A2C | Base | $14.15 \pm 0.12$ | $1.27 \pm 0.01$ | $2.10 \pm 0.02$ | $1.67 \pm 0.01$ | $\mathbf{8.31 \pm 0.03}$ |
| | | FinPILOT | $\mathbf{44.11 \pm 0.84}$ | $\mathbf{2.20 \pm 0.03}$ | $\mathbf{3.51 \pm 0.04}$ | $\mathbf{4.21 \pm 0.16}$ | $9.04 \pm 0.22$ |
| | TD3 | Base | $11.85 \pm 1.84$ | $1.03 \pm 0.18$ | $1.64 \pm 0.30$ | $1.18 \pm 0.25$ | $\mathbf{10.32 \pm 1.77}$ |
| | | FinPILOT | $\mathbf{53.69 \pm 8.19}$ | $\mathbf{2.05 \pm 0.18}$ | $\mathbf{3.56 \pm 0.39}$ | $\mathbf{3.69 \pm 0.60}$ | $12.45 \pm 1.61$ |
| | DDPG | Base | $9.09 \pm 2.25$ | $0.85 \pm 0.19$ | $1.36 \pm 0.32$ | $1.05 \pm 0.30$ | $\mathbf{9.07 \pm 1.49}$ |
| | | FinPILOT | $\mathbf{64.31 \pm 6.98}$ | $\mathbf{2.42 \pm 0.30}$ | $\mathbf{4.55 \pm 0.78}$ | $\mathbf{4.53 \pm 0.96}$ | $11.73 \pm 1.53$ |
| 2024 | PPO | Base | $10.78 \pm 0.13$ | $1.05 \pm 0.01$ | $1.52 \pm 0.02$ | $1.77 \pm 0.03$ | $\mathbf{6.09 \pm 0.07}$ |
| | | FinPILOT | $\mathbf{21.01 \pm 1.12}$ | $\mathbf{1.29 \pm 0.05}$ | $\mathbf{1.83 \pm 0.09}$ | $\mathbf{1.88 \pm 0.09}$ | $10.80 \pm 0.17$ |
| | SAC | Base | $10.95 \pm 0.01$ | $1.06 \pm 0.00$ | $1.54 \pm 0.00$ | $1.77 \pm 0.00$ | $6.15 \pm 0.00$ |
| | | FinPILOT | $\mathbf{11.31 \pm 0.00}$ | $\mathbf{1.09 \pm 0.00}$ | $\mathbf{1.62 \pm 0.00}$ | $\mathbf{2.01 \pm 0.00}$ | $\mathbf{5.60 \pm 0.00}$ |
| | A2C | Base | $10.93 \pm 0.05$ | $\mathbf{1.06 \pm 0.01}$ | $\mathbf{1.53 \pm 0.01}$ | $\mathbf{1.77 \pm 0.02}$ | $\mathbf{6.18 \pm 0.04}$ |
| | | FinPILOT | $\mathbf{12.29 \pm 0.32}$ | $0.98 \pm 0.02$ | $1.42 \pm 0.03$ | $1.37 \pm 0.04$ | $9.07 \pm 0.06$ |
| | TD3 | Base | $11.12 \pm 7.02$ | $1.01 \pm 0.62$ | $1.37 \pm 0.71$ | $1.64 \pm 1.04$ | $\mathbf{6.92 \pm 0.97}$ |
| | | FinPILOT | $\mathbf{26.65 \pm 6.93}$ | $\mathbf{1.41 \pm 0.39}$ | $\mathbf{2.10 \pm 0.66}$ | $\mathbf{2.37 \pm 0.87}$ | $11.80 \pm 3.99$ |
| | DDPG | Base | $10.37 \pm 5.26$ | $0.86 \pm 0.40$ | $1.22 \pm 0.62$ | $1.38 \pm 0.85$ | $\mathbf{8.39 \pm 1.63}$ |
| | | FinPILOT | $\mathbf{23.86 \pm 5.94}$ | $\mathbf{1.41 \pm 0.36}$ | $\mathbf{2.06 \pm 0.70}$ | $\mathbf{2.36 \pm 0.68}$ | $9.97 \pm 2.25$ |
| 2025 | PPO | Base | $15.65 \pm 0.28$ | $1.01 \pm 0.02$ | $1.34 \pm 0.03$ | $1.04 \pm 0.02$ | $15.23 \pm 0.15$ |
| | | FinPILOT | $\mathbf{23.68 \pm 2.23}$ | $\mathbf{1.17 \pm 0.09}$ | $\mathbf{1.61 \pm 0.13}$ | $\mathbf{1.61 \pm 0.11}$ | $\mathbf{14.42 \pm 0.38}$ |
| | SAC | Base | $16.09 \pm 0.08$ | $1.04 \pm 0.00$ | $1.37 \pm 0.00$ | $1.08 \pm 0.00$ | $\mathbf{14.92 \pm 0.01}$ |
| | | FinPILOT | $\mathbf{17.73 \pm 0.06}$ | $\mathbf{1.05 \pm 0.00}$ | $\mathbf{1.44 \pm 0.00}$ | $\mathbf{1.17 \pm 0.00}$ | $15.15 \pm 0.01$ |
| | A2C | Base | $16.11 \pm 0.11$ | $1.04 \pm 0.00$ | $1.37 \pm 0.01$ | $1.08 \pm 0.01$ | $14.96 \pm 0.04$ |
| | | FinPILOT | $\mathbf{25.11 \pm 0.80}$ | $\mathbf{1.18 \pm 0.03}$ | $\mathbf{1.76 \pm 0.05}$ | $\mathbf{1.80 \pm 0.06}$ | $\mathbf{13.67 \pm 0.09}$ |
| | TD3 | Base | $15.82 \pm 4.09$ | $0.95 \pm 0.15$ | $1.28 \pm 0.19$ | $0.93 \pm 0.13$ | $\mathbf{17.09 \pm 2.22}$ |
| | | FinPILOT | $\mathbf{28.56 \pm 12.28}$ | $\mathbf{1.18 \pm 0.40}$ | $\mathbf{1.80 \pm 0.70}$ | $\mathbf{1.59 \pm 0.61}$ | $18.38 \pm 5.07$ |
| | DDPG | Base | $12.94 \pm 4.77$ | $\mathbf{0.85 \pm 0.29}$ | $1.10 \pm 0.34$ | $0.81 \pm 0.26$ | $\mathbf{16.73 \pm 1.96}$ |
| | | FinPILOT | $\mathbf{20.38 \pm 9.89}$ | $0.81 \pm 0.32$ | $\mathbf{1.29 \pm 0.55}$ | $\mathbf{1.30 \pm 0.57}$ | $17.42 \pm 2.61$ |

Table 8: Per-window performance on FX (three rolling test windows, Theta + VAR(1)-ridge forecaster). Values are mean ± one standard deviation across random seeds ($n = 5$ per window cell). Base denotes the frozen pre-trained policy; FinPILOT denotes the same agent under our method. For each window, algorithm, and metric, the better of the two means is shown in **bold** (higher is better for all metrics except MDD, where lower is better); values tied at two decimals are both bolded.

| Window | Algo | Method | Return (%) | Sharpe | Sortino | Calmar | MDD (%) |
|---|---|---|---|---|---|---|---|
| 2023 | PPO | Base | $2.05 \pm 0.13$ | $0.39 \pm 0.02$ | $0.65 \pm 0.04$ | $0.40 \pm 0.03$ | $5.51 \pm 0.22$ |
| | | FinPILOT | $\mathbf{2.33 \pm 0.97}$ | $\mathbf{0.54 \pm 0.21}$ | $\mathbf{0.92 \pm 0.33}$ | $\mathbf{0.45 \pm 0.20}$ | $\mathbf{5.42 \pm 0.52}$ |
| | SAC | Base | $2.32 \pm 0.02$ | $0.45 \pm 0.00$ | $0.74 \pm 0.01$ | $\mathbf{0.46 \pm 0.00}$ | $5.35 \pm 0.01$ |
| | | FinPILOT | $\mathbf{2.33 \pm 0.02}$ | $\mathbf{0.49 \pm 0.00}$ | $\mathbf{0.82 \pm 0.01}$ | $\mathbf{0.46 \pm 0.00}$ | $\mathbf{5.28 \pm 0.01}$ |
| | A2C | Base | $2.31 \pm 0.03$ | $0.45 \pm 0.01$ | $0.74 \pm 0.01$ | $0.46 \pm 0.01$ | $5.35 \pm 0.01$ |
| | | FinPILOT | $\mathbf{2.78 \pm 0.03}$ | $\mathbf{0.61 \pm 0.01}$ | $\mathbf{1.06 \pm 0.02}$ | $\mathbf{0.54 \pm 0.01}$ | $\mathbf{5.27 \pm 0.05}$ |
| | TD3 | Base | $2.63 \pm 0.84$ | $0.51 \pm 0.15$ | $0.82 \pm 0.23$ | $0.50 \pm 0.19$ | $\mathbf{5.67 \pm 0.63}$ |
| | | FinPILOT | $\mathbf{4.76 \pm 1.78}$ | $\mathbf{0.90 \pm 0.27}$ | $\mathbf{1.49 \pm 0.47}$ | $\mathbf{0.75 \pm 0.26}$ | $6.45 \pm 1.20$ |
| | DDPG | Base | $2.62 \pm 1.77$ | $0.51 \pm 0.31$ | $0.81 \pm 0.51$ | $0.50 \pm 0.30$ | $\mathbf{5.70 \pm 0.83}$ |
| | | FinPILOT | $\mathbf{4.96 \pm 2.60}$ | $\mathbf{1.00 \pm 0.44}$ | $\mathbf{1.63 \pm 0.80}$ | $\mathbf{0.79 \pm 0.38}$ | $6.31 \pm 0.83$ |
| 2024 | PPO | Base | $-5.82 \pm 0.10$ | $-1.36 \pm 0.01$ | $-1.87 \pm 0.02$ | $\mathbf{-0.90 \pm 0.01}$ | $6.53 \pm 0.10$ |
| | | FinPILOT | $\mathbf{-4.60 \pm 0.24}$ | $\mathbf{-1.24 \pm 0.05}$ | $\mathbf{-1.67 \pm 0.08}$ | $-0.95 \pm 0.04$ | $\mathbf{4.85 \pm 0.34}$ |
| | SAC | Base | $-5.74 \pm 0.07$ | $\mathbf{-1.35 \pm 0.01}$ | $\mathbf{-1.87 \pm 0.02}$ | $\mathbf{-0.90 \pm 0.01}$ | $6.49 \pm 0.04$ |
| | | FinPILOT | $\mathbf{-5.53 \pm 0.05}$ | $-1.42 \pm 0.01$ | $-1.89 \pm 0.01$ | $-1.00 \pm 0.01$ | $\mathbf{5.60 \pm 0.01}$ |
| | A2C | Base | $-5.78 \pm 0.00$ | $-1.36 \pm 0.00$ | $-1.88 \pm 0.00$ | $-0.90 \pm 0.00$ | $6.51 \pm 0.01$ |
| | | FinPILOT | $\mathbf{-3.79 \pm 0.24}$ | $\mathbf{-1.05 \pm 0.05}$ | $\mathbf{-1.43 \pm 0.07}$ | $\mathbf{-0.84 \pm 0.02}$ | $\mathbf{4.51 \pm 0.19}$ |
| | TD3 | Base | $-7.36 \pm 1.37$ | $-1.56 \pm 0.24$ | $-2.24 \pm 0.27$ | $-0.99 \pm 0.05$ | $7.43 \pm 1.21$ |
| | | FinPILOT | $\mathbf{-3.23 \pm 1.51}$ | $\mathbf{-0.83 \pm 0.33}$ | $\mathbf{-1.18 \pm 0.39}$ | $\mathbf{-0.67 \pm 0.14}$ | $\mathbf{4.55 \pm 1.44}$ |
| | DDPG | Base | $-6.78 \pm 1.31$ | $-1.62 \pm 0.28$ | $-2.33 \pm 0.49$ | $-1.01 \pm 0.03$ | $6.68 \pm 1.20$ |
| | | FinPILOT | $\mathbf{-4.36 \pm 1.58}$ | $\mathbf{-1.25 \pm 0.48}$ | $\mathbf{-1.61 \pm 0.48}$ | $\mathbf{-0.87 \pm 0.15}$ | $\mathbf{4.93 \pm 1.48}$ |
| 2025 | PPO | Base | $\mathbf{7.00 \pm 0.26}$ | $1.40 \pm 0.04$ | $2.41 \pm 0.08$ | $2.70 \pm 0.11$ | $2.55 \pm 0.04$ |
| | | FinPILOT | $6.14 \pm 0.78$ | $\mathbf{1.54 \pm 0.15}$ | $\mathbf{3.05 \pm 0.29}$ | $\mathbf{3.31 \pm 0.85}$ | $\mathbf{1.91 \pm 0.45}$ |
| | SAC | Base | $7.02 \pm 0.07$ | $1.39 \pm 0.00$ | $2.39 \pm 0.01$ | $2.70 \pm 0.01$ | $2.56 \pm 0.03$ |
| | | FinPILOT | $\mathbf{7.06 \pm 0.04}$ | $\mathbf{1.54 \pm 0.01}$ | $\mathbf{2.80 \pm 0.02}$ | $\mathbf{3.07 \pm 0.01}$ | $\mathbf{2.25 \pm 0.00}$ |
| | A2C | Base | $\mathbf{7.06 \pm 0.02}$ | $1.39 \pm 0.00$ | $2.39 \pm 0.01$ | $\mathbf{2.71 \pm 0.01}$ | $2.57 \pm 0.01$ |
| | | FinPILOT | $5.34 \pm 0.08$ | $\mathbf{1.43 \pm 0.01}$ | $\mathbf{2.93 \pm 0.02}$ | $2.66 \pm 0.04$ | $\mathbf{1.98 \pm 0.01}$ |
| | TD3 | Base | $6.35 \pm 1.05$ | $1.33 \pm 0.19$ | $2.32 \pm 0.29$ | $2.64 \pm 0.74$ | $2.47 \pm 0.47$ |
| | | FinPILOT | $\mathbf{8.74 \pm 0.86}$ | $\mathbf{1.53 \pm 0.09}$ | $\mathbf{2.86 \pm 0.18}$ | $\mathbf{3.55 \pm 0.18}$ | $\mathbf{2.40 \pm 0.22}$ |
| | DDPG | Base | $3.52 \pm 1.85$ | $0.79 \pm 0.39$ | $1.35 \pm 0.67$ | $\mathbf{1.33 \pm 0.97}$ | $\mathbf{3.18 \pm 1.10}$ |
| | | FinPILOT | $\mathbf{3.84 \pm 0.91}$ | $\mathbf{0.82 \pm 0.17}$ | $\mathbf{1.51 \pm 0.28}$ | $1.07 \pm 0.39$ | $3.76 \pm 0.62$ |

### D.2 COMPARISON WITH NON-RL BASELINES

Tables 9 and 10 report per-window results for the non-RL baselines, underlying the three-window averages in Table 3; their construction and hyperparameter selection are described in Appendix C.4. On DJ30, the best FinPILOT-adapted policy exceeds every baseline in total return and Sortino ratio in all three windows, while Boyd SPO attains the highest Sharpe ratio in 2024 and Boyd MPO the lowest drawdown in 2023. Individual adapted policies do not always beat the passive allocators; for example, FinPILOT-SAC and FinPILOT-A2C trail buy-and-hold in 2024 return. On FX, the Boyd optimizers remain essentially in cash in every window, so their near-zero returns and drawdowns reflect inactivity. Among portfolios exposed to the market, a FinPILOT-adapted policy attains the highest return in 2023 and 2025, the smallest loss in 2024 ($-3.23\%$ for TD3 vs. $-5.91\%$ for buy-and-hold), and the lowest drawdown in all three windows.

Table 9: Per-window comparison of FinPILOT with non-RL baselines on DJ30 (XGBoost forecaster). FinPILOT values are mean $\pm$ one standard deviation across random seeds ($n = 5$ per window cell); the baselines are deterministic. SPO and MPO use the same forecast as FinPILOT, with hyperparameters selected on the years preceding each test year (Appendix C.4); the other baselines use no forecast. Best mean per window and column in **bold** (lower is better for MDD); values tied at two decimals are all bolded.

| Window | Method | Return (%) | Sharpe | Sortino | Calmar | MDD (%) |
|---|---|---|---|---|---|---|
| 2023 | Buy-and-hold | 15.61 | 1.33 | 2.22 | 1.72 | 8.84 |
| | 1/N (daily rebalanced) | 14.79 | 1.28 | 2.12 | 1.68 | 8.59 |
| | SPO ($H$=1) | $-4.94$ | $-0.37$ | $-0.42$ | $-0.28$ | 15.40 |
| | MPO ($H$=50) | 2.24 | 0.33 | 0.31 | 0.35 | **7.30** |
| | FinPILOT-PPO | $26.16 \pm 0.43$ | $1.50 \pm 0.02$ | $2.34 \pm 0.04$ | $2.03 \pm 0.03$ | $12.11 \pm 0.00$ |
| | FinPILOT-SAC | $18.82 \pm 0.01$ | $1.51 \pm 0.00$ | $2.61 \pm 0.00$ | $2.01 \pm 0.00$ | $8.94 \pm 0.00$ |
| | FinPILOT-A2C | $44.11 \pm 0.84$ | $2.20 \pm 0.03$ | $3.51 \pm 0.04$ | $4.21 \pm 0.16$ | $9.04 \pm 0.22$ |
| | FinPILOT-TD3 | $53.69 \pm 8.19$ | $2.05 \pm 0.18$ | $3.56 \pm 0.39$ | $3.69 \pm 0.60$ | $12.45 \pm 1.61$ |
| | FinPILOT-DDPG | $\mathbf{64.31 \pm 6.98}$ | $\mathbf{2.42 \pm 0.30}$ | $\mathbf{4.55 \pm 0.78}$ | $\mathbf{4.53 \pm 0.96}$ | $11.73 \pm 1.53$ |
| 2024 | Buy-and-hold | 12.62 | 1.18 | 1.75 | 2.14 | 5.81 |
| | 1/N (daily rebalanced) | 11.34 | 1.06 | 1.54 | 1.78 | 6.37 |
| | SPO ($H$=1) | 20.54 | **1.42** | 1.64 | 2.27 | 8.66 |
| | MPO ($H$=50) | 4.42 | 0.30 | 0.37 | 0.37 | 20.47 |
| | FinPILOT-PPO | $21.01 \pm 1.12$ | $1.29 \pm 0.05$ | $1.83 \pm 0.09$ | $1.88 \pm 0.09$ | $10.80 \pm 0.17$ |
| | FinPILOT-SAC | $11.31 \pm 0.00$ | $1.09 \pm 0.00$ | $1.62 \pm 0.00$ | $2.01 \pm 0.00$ | $\mathbf{5.60 \pm 0.00}$ |
| | FinPILOT-A2C | $12.29 \pm 0.32$ | $0.98 \pm 0.02$ | $1.42 \pm 0.03$ | $1.37 \pm 0.04$ | $9.07 \pm 0.06$ |
| | FinPILOT-TD3 | $\mathbf{26.65 \pm 6.93}$ | $1.41 \pm 0.39$ | $\mathbf{2.10 \pm 0.66}$ | $\mathbf{2.37 \pm 0.87}$ | $11.80 \pm 3.99$ |
| | FinPILOT-DDPG | $23.86 \pm 5.94$ | $1.41 \pm 0.36$ | $2.06 \pm 0.70$ | $2.36 \pm 0.68$ | $9.97 \pm 2.25$ |
| 2025 | Buy-and-hold | 15.86 | 1.01 | 1.34 | 1.04 | 15.43 |
| | 1/N (daily rebalanced) | 16.64 | 1.04 | 1.37 | 1.08 | 15.42 |
| | SPO ($H$=1) | 11.26 | 0.49 | 0.68 | 0.55 | 28.81 |
| | MPO ($H$=50) | 12.08 | 0.49 | 0.78 | 0.65 | 29.66 |
| | FinPILOT-PPO | $23.68 \pm 2.23$ | $1.17 \pm 0.09$ | $1.61 \pm 0.13$ | $1.61 \pm 0.11$ | $14.42 \pm 0.38$ |
| | FinPILOT-SAC | $17.73 \pm 0.06$ | $1.05 \pm 0.00$ | $1.44 \pm 0.00$ | $1.17 \pm 0.00$ | $15.15 \pm 0.01$ |
| | FinPILOT-A2C | $25.11 \pm 0.80$ | $\mathbf{1.18 \pm 0.03}$ | $1.76 \pm 0.05$ | $\mathbf{1.80 \pm 0.06}$ | $\mathbf{13.67 \pm 0.09}$ |
| | FinPILOT-TD3 | $\mathbf{28.56 \pm 12.28}$ | $\mathbf{1.18 \pm 0.40}$ | $\mathbf{1.80 \pm 0.70}$ | $1.59 \pm 0.61$ | $18.38 \pm 5.07$ |
| | FinPILOT-DDPG | $20.38 \pm 9.89$ | $0.81 \pm 0.32$ | $1.29 \pm 0.55$ | $1.30 \pm 0.57$ | $17.42 \pm 2.61$ |

Table 10: Per-window comparison of FinPILOT with non-RL baselines on FX (Theta + VAR(1)-ridge forecaster). FinPILOT values are mean ± one standard deviation across random seeds ($n = 5$ per window cell); the baselines are deterministic. SPO and MPO use the same forecast as FinPILOT, with hyperparameters selected on the years preceding each test year (Appendix C.4); the other baselines use no forecast. †SPO and MPO remain essentially in cash, losing less than \$1 on \$100,000 in every window; their returns and drawdowns are below 0.01% in magnitude. Their Calmar and MDD are excluded when marking the best value, and their Sharpe, Sortino, and Calmar ratios, computed from these near-zero returns, are numerically unstable. Best mean per window and column in **bold** (lower is better for MDD); values tied at two decimals are all bolded.

| Window | Method | Return (%) | Sharpe | Sortino | Calmar | MDD (%) |
|---|---|---|---|---|---|---|
| 2023 | Buy-and-hold | 2.39 | 0.45 | 0.74 | 0.45 | 5.59 |
| | 1/N (daily rebalanced) | 2.42 | 0.45 | 0.74 | 0.46 | 5.58 |
| | SPO ($H$=1) | 0.00† | −0.01 | −0.02 | −0.01† | <0.01† |
| | MPO ($H$=50) | 0.00† | −1.75 | −2.07 | −0.88† | <0.01† |
| | FinPILOT-PPO | 2.33 ± 0.97 | 0.54 ± 0.21 | 0.92 ± 0.33 | 0.45 ± 0.20 | 5.42 ± 0.52 |
| | FinPILOT-SAC | 2.33 ± 0.02 | 0.49 ± 0.00 | 0.82 ± 0.01 | 0.46 ± 0.00 | 5.28 ± 0.01 |
| | FinPILOT-A2C | 2.78 ± 0.03 | 0.61 ± 0.01 | 1.06 ± 0.02 | 0.54 ± 0.01 | **5.27 ± 0.05** |
| | FinPILOT-TD3 | 4.76 ± 1.78 | 0.90 ± 0.27 | 1.49 ± 0.47 | 0.75 ± 0.26 | 6.45 ± 1.20 |
| | FinPILOT-DDPG | **4.96 ± 2.60** | **1.00 ± 0.44** | **1.63 ± 0.80** | **0.79 ± 0.38** | 6.31 ± 0.83 |
| 2024 | Buy-and-hold | −5.91 | −1.34 | −1.83 | −0.89 | 6.73 |
| | 1/N (daily rebalanced) | −6.03 | −1.36 | −1.87 | −0.90 | 6.80 |
| | SPO ($H$=1) | **0.00†** | −1.56 | −1.69 | −0.92† | <0.01† |
| | MPO ($H$=50) | **0.00†** | −4.29 | −4.70 | −1.00† | <0.01† |
| | FinPILOT-PPO | −4.60 ± 0.24 | −1.24 ± 0.05 | −1.67 ± 0.08 | −0.95 ± 0.04 | 4.85 ± 0.34 |
| | FinPILOT-SAC | −5.53 ± 0.05 | −1.42 ± 0.01 | −1.89 ± 0.01 | −1.00 ± 0.01 | 5.60 ± 0.01 |
| | FinPILOT-A2C | −3.79 ± 0.24 | −1.05 ± 0.05 | −1.43 ± 0.07 | −0.84 ± 0.02 | **4.51 ± 0.19** |
| | FinPILOT-TD3 | −3.23 ± 1.51 | **−0.83 ± 0.33** | **−1.18 ± 0.39** | **−0.67 ± 0.14** | 4.55 ± 1.44 |
| | FinPILOT-DDPG | −4.36 ± 1.58 | −1.25 ± 0.48 | −1.61 ± 0.48 | −0.87 ± 0.15 | 4.93 ± 1.48 |
| 2025 | Buy-and-hold | 7.46 | 1.39 | 2.39 | 2.72 | 2.70 |
| | 1/N (daily rebalanced) | 7.38 | 1.39 | 2.40 | 2.70 | 2.69 |
| | SPO ($H$=1) | 0.00† | −0.57 | −0.71 | −0.64† | <0.01† |
| | MPO ($H$=50) | 0.00† | −2.89 | −3.77 | −0.98† | <0.01† |
| | FinPILOT-PPO | 6.14 ± 0.78 | **1.54 ± 0.15** | **3.05 ± 0.29** | 3.31 ± 0.85 | **1.91 ± 0.45** |
| | FinPILOT-SAC | 7.06 ± 0.04 | **1.54 ± 0.01** | 2.80 ± 0.02 | 3.07 ± 0.01 | 2.25 ± 0.00 |
| | FinPILOT-A2C | 5.34 ± 0.08 | 1.43 ± 0.01 | 2.93 ± 0.02 | 2.66 ± 0.04 | 1.98 ± 0.01 |
| | FinPILOT-TD3 | **8.74 ± 0.86** | 1.53 ± 0.09 | 2.86 ± 0.18 | **3.55 ± 0.18** | 2.40 ± 0.22 |
| | FinPILOT-DDPG | 3.84 ± 0.91 | 0.82 ± 0.17 | 1.51 ± 0.28 | 1.07 ± 0.39 | 3.76 ± 0.62 |

### D.3 Controlled forecast-skill results

Tables 11–16 report per-window results for every target $R^2_\star$, algorithm, and market, underlying Figure 4. Controlled forecasts are constructed as in Section 4.4, with $R^2_{t,i}$ defined in Appendix C.3. The Base row is the frozen pre-trained policy, and the Original forecast row uses the learned forecaster.

Table 11: FinPILOT performance with controlled forecasts on DJ30, test year 2023, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | 13.84 ± 0.39 | 1.25 ± 0.02 | 2.07 ± 0.04 | 1.64 ± 0.04 | 8.27 ± 0.05 |
| | Original forecast | 26.16 ± 0.43 | 1.50 ± 0.02 | 2.34 ± 0.04 | 2.03 ± 0.03 | 12.11 ± 0.00 |
| | $R^2_\star = 0.005$ | 25.44 ± 0.38 | 1.52 ± 0.02 | 2.44 ± 0.03 | 1.95 ± 0.03 | 12.27 ± 0.07 |
| | $R^2_\star = 0.01$ | 25.74 ± 0.36 | 1.54 ± 0.02 | 2.46 ± 0.03 | 1.97 ± 0.03 | 12.29 ± 0.06 |
| | $R^2_\star = 0.025$ | 26.73 ± 0.34 | 1.58 ± 0.02 | 2.55 ± 0.03 | 2.02 ± 0.03 | 12.35 ± 0.04 |
| | $R^2_\star = 0.05$ | 28.67 ± 0.28 | 1.67 ± 0.01 | 2.74 ± 0.03 | 2.13 ± 0.02 | 12.41 ± 0.02 |
| | $R^2_\star = 0.1$ | 32.11 ± 0.36 | 1.83 ± 0.02 | 3.00 ± 0.03 | 2.34 ± 0.02 | 12.46 ± 0.03 |
| | $R^2_\star = 0.2$ | 51.38 ± 0.63 | 2.66 ± 0.02 | 4.73 ± 0.05 | 3.46 ± 0.03 | 12.36 ± 0.10 |
| | $R^2_\star = 0.4$ | 93.19 ± 0.91 | 3.66 ± 0.02 | 7.32 ± 0.08 | 6.70 ± 0.21 | 10.11 ± 0.25 |
| | $R^2_\star = 0.8$ | 128.31 ± 1.00 | 4.49 ± 0.02 | 8.94 ± 0.06 | 9.68 ± 0.10 | 8.73 ± 0.11 |
| **SAC** | Base | 14.25 ± 0.01 | 1.28 ± 0.00 | 2.12 ± 0.00 | 1.68 ± 0.00 | 8.30 ± 0.00 |
| | Original forecast | 18.82 ± 0.01 | 1.51 ± 0.00 | 2.61 ± 0.00 | 2.01 ± 0.00 | 8.94 ± 0.00 |
| | $R^2_\star = 0.005$ | 17.85 ± 0.01 | 1.44 ± 0.00 | 2.47 ± 0.00 | 1.86 ± 0.00 | 9.24 ± 0.00 |
| | $R^2_\star = 0.01$ | 18.05 ± 0.01 | 1.45 ± 0.00 | 2.49 ± 0.00 | 1.88 ± 0.00 | 9.22 ± 0.00 |
| | $R^2_\star = 0.025$ | 18.64 ± 0.01 | 1.49 ± 0.00 | 2.57 ± 0.00 | 1.94 ± 0.00 | 9.17 ± 0.00 |
| | $R^2_\star = 0.05$ | 19.56 ± 0.01 | 1.56 ± 0.00 | 2.67 ± 0.00 | 2.05 ± 0.00 | 9.08 ± 0.00 |
| | $R^2_\star = 0.1$ | 21.65 ± 0.01 | 1.70 ± 0.00 | 2.93 ± 0.00 | 2.29 ± 0.00 | 8.88 ± 0.00 |
| | $R^2_\star = 0.2$ | 26.78 ± 0.02 | 2.04 ± 0.00 | 3.54 ± 0.00 | 2.96 ± 0.00 | 8.27 ± 0.00 |
| | $R^2_\star = 0.4$ | 35.61 ± 0.01 | 2.58 ± 0.00 | 4.55 ± 0.00 | 4.16 ± 0.00 | 7.50 ± 0.00 |
| | $R^2_\star = 0.8$ | 42.19 ± 0.01 | 3.06 ± 0.00 | 5.49 ± 0.00 | 6.10 ± 0.00 | 5.89 ± 0.00 |
| **A2C** | Base | 14.15 ± 0.12 | 1.27 ± 0.01 | 2.10 ± 0.02 | 1.67 ± 0.01 | 8.31 ± 0.03 |
| | Original forecast | 44.11 ± 0.84 | 2.20 ± 0.03 | 3.51 ± 0.04 | 4.21 ± 0.16 | 9.04 ± 0.22 |
| | $R^2_\star = 0.005$ | 39.71 ± 0.95 | 1.99 ± 0.03 | 3.07 ± 0.05 | 3.63 ± 0.17 | 9.67 ± 0.26 |
| | $R^2_\star = 0.01$ | 43.23 ± 0.98 | 2.09 ± 0.03 | 3.27 ± 0.06 | 4.23 ± 0.22 | 8.90 ± 0.31 |
| | $R^2_\star = 0.025$ | 52.50 ± 0.24 | 2.31 ± 0.01 | 3.81 ± 0.02 | 5.60 ± 0.03 | 7.86 ± 0.02 |
| | $R^2_\star = 0.05$ | 56.62 ± 0.28 | 2.34 ± 0.01 | 4.10 ± 0.01 | 5.99 ± 0.02 | 7.84 ± 0.00 |
| | $R^2_\star = 0.1$ | 70.14 ± 0.91 | 2.57 ± 0.04 | 4.77 ± 0.09 | 6.10 ± 0.43 | 9.14 ± 0.58 |
| | $R^2_\star = 0.2$ | 79.49 ± 0.29 | 2.89 ± 0.01 | 5.25 ± 0.02 | 6.89 ± 0.12 | 8.82 ± 0.13 |
| | $R^2_\star = 0.4$ | 119.15 ± 0.71 | 4.07 ± 0.02 | 7.40 ± 0.02 | 10.20 ± 0.04 | 7.89 ± 0.00 |
| | $R^2_\star = 0.8$ | 145.02 ± 0.30 | 4.25 ± 0.01 | 7.91 ± 0.02 | 9.64 ± 0.04 | 9.56 ± 0.03 |
| **TD3** | Base | 11.85 ± 1.84 | 1.03 ± 0.18 | 1.64 ± 0.30 | 1.18 ± 0.25 | 10.32 ± 1.77 |
| | Original forecast | 53.69 ± 8.19 | 2.05 ± 0.18 | 3.56 ± 0.39 | 3.69 ± 0.60 | 12.45 ± 1.61 |
| | $R^2_\star = 0.005$ | 41.57 ± 17.56 | 1.66 ± 0.49 | 2.81 ± 1.12 | 2.76 ± 0.96 | 13.45 ± 0.92 |
| | $R^2_\star = 0.01$ | 49.98 ± 12.59 | 1.90 ± 0.29 | 3.15 ± 0.88 | 3.55 ± 0.58 | 12.10 ± 1.13 |
| | $R^2_\star = 0.025$ | 50.96 ± 14.85 | 1.97 ± 0.40 | 3.42 ± 0.93 | 3.90 ± 1.16 | 11.51 ± 2.29 |
| | $R^2_\star = 0.05$ | 58.75 ± 13.94 | 2.21 ± 0.36 | 3.95 ± 0.76 | 3.89 ± 0.93 | 12.66 ± 1.68 |
| | $R^2_\star = 0.1$ | 70.21 ± 11.19 | 2.54 ± 0.33 | 4.59 ± 0.74 | 5.16 ± 1.37 | 11.14 ± 1.99 |
| | $R^2_\star = 0.2$ | 69.98 ± 12.03 | 2.42 ± 0.25 | 4.45 ± 0.55 | 4.75 ± 1.61 | 12.24 ± 2.08 |
| | $R^2_\star = 0.4$ | 103.54 ± 19.85 | 3.34 ± 0.31 | 6.49 ± 0.82 | 9.20 ± 2.48 | 8.25 ± 1.48 |
| | $R^2_\star = 0.8$ | 548.88 ± 396.61 | 7.07 ± 1.20 | 18.05 ± 5.49 | 33.08 ± 17.86 | 5.89 ± 1.72 |
| **DDPG** | Base | 9.09 ± 2.25 | 0.85 ± 0.19 | 1.36 ± 0.32 | 1.05 ± 0.30 | 9.07 ± 1.49 |
| | Original forecast | 64.31 ± 6.98 | 2.42 ± 0.30 | 4.55 ± 0.78 | 4.53 ± 0.96 | 11.73 ± 1.53 |
| | $R^2_\star = 0.005$ | 55.04 ± 15.26 | 2.17 ± 0.32 | 3.88 ± 0.71 | 3.84 ± 0.93 | 12.12 ± 2.03 |
| | $R^2_\star = 0.01$ | 59.53 ± 8.64 | 2.25 ± 0.25 | 4.20 ± 0.58 | 3.72 ± 0.19 | 13.17 ± 1.41 |
| | $R^2_\star = 0.025$ | 64.80 ± 10.41 | 2.42 ± 0.24 | 4.41 ± 0.63 | 4.48 ± 0.94 | 11.94 ± 2.12 |
| | $R^2_\star = 0.05$ | 63.00 ± 13.43 | 2.27 ± 0.31 | 4.06 ± 0.78 | 3.94 ± 0.99 | 13.28 ± 1.46 |
| | $R^2_\star = 0.1$ | 74.40 ± 2.61 | 2.60 ± 0.15 | 4.77 ± 0.28 | 4.51 ± 0.61 | 13.06 ± 1.54 |
| | $R^2_\star = 0.2$ | 82.85 ± 13.71 | 2.77 ± 0.35 | 5.16 ± 0.81 | 6.22 ± 1.82 | 10.59 ± 2.46 |
| | $R^2_\star = 0.4$ | 111.89 ± 31.58 | 3.33 ± 0.77 | 6.59 ± 1.63 | 8.59 ± 3.72 | 9.90 ± 2.83 |
| | $R^2_\star = 0.8$ | 993.53 ± 601.88 | 8.23 ± 1.78 | 20.18 ± 5.22 | 33.67 ± 22.23 | 7.97 ± 3.17 |

Table 12: FinPILOT performance with controlled forecasts on DJ30, test year 2024, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | 10.78 ± 0.13 | 1.05 ± 0.01 | 1.52 ± 0.02 | 1.77 ± 0.03 | 6.09 ± 0.07 |
| | Original forecast | 21.01 ± 1.12 | 1.29 ± 0.05 | 1.83 ± 0.09 | 1.88 ± 0.09 | 10.80 ± 0.17 |
| | $R^2_\star = 0.005$ | 16.07 ± 0.25 | 1.04 ± 0.01 | 1.62 ± 0.02 | 1.74 ± 0.02 | 9.26 ± 0.00 |
| | $R^2_\star = 0.01$ | 16.95 ± 0.48 | 1.09 ± 0.03 | 1.69 ± 0.04 | 1.82 ± 0.04 | 9.27 ± 0.00 |
| | $R^2_\star = 0.025$ | 20.43 ± 0.11 | 1.28 ± 0.01 | 1.99 ± 0.01 | 2.14 ± 0.01 | 9.27 ± 0.00 |
| | $R^2_\star = 0.05$ | 24.75 ± 0.19 | 1.51 ± 0.01 | 2.36 ± 0.02 | 2.53 ± 0.02 | 9.20 ± 0.05 |
| | $R^2_\star = 0.1$ | 45.10 ± 0.26 | 2.37 ± 0.02 | 4.13 ± 0.02 | 4.21 ± 0.02 | 9.16 ± 0.04 |
| | $R^2_\star = 0.2$ | 62.02 ± 0.19 | 2.78 ± 0.01 | 5.03 ± 0.02 | 5.57 ± 0.08 | 8.97 ± 0.12 |
| | $R^2_\star = 0.4$ | 101.24 ± 0.34 | 4.28 ± 0.00 | 7.24 ± 0.03 | 11.85 ± 0.03 | 6.03 ± 0.00 |
| | $R^2_\star = 0.8$ | 128.77 ± 0.31 | 5.10 ± 0.01 | 9.34 ± 0.01 | 17.28 ± 0.03 | 4.87 ± 0.00 |
| **SAC** | Base | 10.95 ± 0.01 | 1.06 ± 0.00 | 1.54 ± 0.00 | 1.77 ± 0.00 | 6.15 ± 0.00 |
| | Original forecast | 11.31 ± 0.00 | 1.09 ± 0.00 | 1.62 ± 0.00 | 2.01 ± 0.00 | 5.60 ± 0.00 |
| | $R^2_\star = 0.005$ | 11.67 ± 0.00 | 1.11 ± 0.00 | 1.67 ± 0.00 | 2.03 ± 0.00 | 5.71 ± 0.00 |
| | $R^2_\star = 0.01$ | 12.02 ± 0.00 | 1.14 ± 0.00 | 1.71 ± 0.00 | 2.09 ± 0.00 | 5.71 ± 0.00 |
| | $R^2_\star = 0.025$ | 13.08 ± 0.00 | 1.22 ± 0.00 | 1.85 ± 0.00 | 2.26 ± 0.00 | 5.69 ± 0.00 |
| | $R^2_\star = 0.05$ | 14.83 ± 0.01 | 1.36 ± 0.00 | 2.07 ± 0.00 | 2.53 ± 0.00 | 5.69 ± 0.00 |
| | $R^2_\star = 0.1$ | 18.22 ± 0.00 | 1.62 ± 0.00 | 2.48 ± 0.00 | 3.04 ± 0.00 | 5.70 ± 0.00 |
| | $R^2_\star = 0.2$ | 24.22 ± 0.01 | 2.07 ± 0.00 | 3.22 ± 0.00 | 4.07 ± 0.00 | 5.48 ± 0.00 |
| | $R^2_\star = 0.4$ | 31.77 ± 0.01 | 2.55 ± 0.00 | 4.02 ± 0.00 | 5.80 ± 0.00 | 4.87 ± 0.00 |
| | $R^2_\star = 0.8$ | 36.50 ± 0.01 | 2.80 ± 0.00 | 4.33 ± 0.00 | 6.92 ± 0.00 | 4.59 ± 0.00 |
| **A2C** | Base | 10.93 ± 0.05 | 1.06 ± 0.01 | 1.53 ± 0.01 | 1.77 ± 0.02 | 6.18 ± 0.04 |
| | Original forecast | 12.29 ± 0.32 | 0.98 ± 0.02 | 1.42 ± 0.03 | 1.37 ± 0.04 | 9.07 ± 0.06 |
| | $R^2_\star = 0.005$ | 15.54 ± 1.07 | 1.19 ± 0.06 | 1.76 ± 0.10 | 2.01 ± 0.14 | 7.59 ± 0.05 |
| | $R^2_\star = 0.01$ | 17.92 ± 0.39 | 1.32 ± 0.02 | 1.99 ± 0.04 | 2.34 ± 0.06 | 7.43 ± 0.05 |
| | $R^2_\star = 0.025$ | 20.51 ± 0.44 | 1.44 ± 0.03 | 2.24 ± 0.04 | 2.83 ± 0.08 | 6.91 ± 0.10 |
| | $R^2_\star = 0.05$ | 27.09 ± 0.47 | 1.78 ± 0.04 | 2.86 ± 0.05 | 3.41 ± 0.12 | 7.33 ± 0.24 |
| | $R^2_\star = 0.1$ | 32.74 ± 0.43 | 1.88 ± 0.03 | 3.13 ± 0.05 | 3.33 ± 0.04 | 8.87 ± 0.01 |
| | $R^2_\star = 0.2$ | 51.26 ± 0.45 | 2.70 ± 0.02 | 4.63 ± 0.03 | 5.23 ± 0.06 | 8.15 ± 0.03 |
| | $R^2_\star = 0.4$ | 98.09 ± 0.48 | 4.37 ± 0.02 | 8.33 ± 0.04 | 14.21 ± 0.32 | 4.91 ± 0.09 |
| | $R^2_\star = 0.8$ | 107.48 ± 0.28 | 4.34 ± 0.01 | 7.66 ± 0.02 | 13.04 ± 0.03 | 5.72 ± 0.01 |
| **TD3** | Base | 11.12 ± 7.02 | 1.01 ± 0.62 | 1.37 ± 0.71 | 1.64 ± 1.04 | 6.92 ± 0.97 |
| | Original forecast | 26.65 ± 6.93 | 1.41 ± 0.39 | 2.10 ± 0.66 | 2.37 ± 0.87 | 11.80 ± 3.99 |
| | $R^2_\star = 0.005$ | 10.88 ± 11.59 | 0.69 ± 0.69 | 0.97 ± 0.97 | 1.19 ± 1.39 | 12.57 ± 3.11 |
| | $R^2_\star = 0.01$ | 14.25 ± 8.97 | 0.87 ± 0.52 | 1.23 ± 0.76 | 1.37 ± 0.99 | 12.75 ± 3.83 |
| | $R^2_\star = 0.025$ | 25.09 ± 6.45 | 1.50 ± 0.35 | 2.29 ± 0.69 | 3.06 ± 1.51 | 9.06 ± 3.69 |
| | $R^2_\star = 0.05$ | 34.36 ± 7.55 | 1.75 ± 0.42 | 2.69 ± 0.88 | 4.11 ± 1.91 | 8.72 ± 3.83 |
| | $R^2_\star = 0.1$ | 49.20 ± 13.67 | 2.26 ± 0.50 | 3.57 ± 1.10 | 4.82 ± 1.36 | 8.80 ± 1.53 |
| | $R^2_\star = 0.2$ | 103.32 ± 39.67 | 3.77 ± 0.90 | 6.79 ± 1.97 | 10.54 ± 3.42 | 7.08 ± 1.59 |
| | $R^2_\star = 0.4$ | 123.49 ± 59.28 | 4.22 ± 0.85 | 7.53 ± 2.12 | 11.98 ± 4.53 | 6.72 ± 1.42 |
| | $R^2_\star = 0.8$ | 577.55 ± 306.04 | 7.47 ± 1.87 | 17.74 ± 7.87 | 43.59 ± 21.61 | 4.52 ± 1.23 |
| **DDPG** | Base | 10.37 ± 5.26 | 0.86 ± 0.40 | 1.22 ± 0.62 | 1.38 ± 0.85 | 8.39 ± 1.63 |
| | Original forecast | 23.86 ± 5.94 | 1.41 ± 0.36 | 2.06 ± 0.70 | 2.36 ± 0.68 | 9.97 ± 2.25 |
| | $R^2_\star = 0.005$ | 8.68 ± 8.52 | 0.54 ± 0.35 | 0.74 ± 0.50 | 0.82 ± 0.73 | 14.20 ± 4.85 |
| | $R^2_\star = 0.01$ | 6.87 ± 14.42 | 0.43 ± 0.74 | 0.70 ± 1.19 | 0.65 ± 1.06 | 15.44 ± 4.61 |
| | $R^2_\star = 0.025$ | 22.69 ± 10.79 | 1.32 ± 0.63 | 1.97 ± 1.10 | 2.49 ± 1.32 | 9.20 ± 1.29 |
| | $R^2_\star = 0.05$ | 30.27 ± 3.72 | 1.62 ± 0.23 | 2.44 ± 0.29 | 2.68 ± 0.80 | 11.16 ± 3.15 |
| | $R^2_\star = 0.1$ | 36.10 ± 7.94 | 1.97 ± 0.29 | 2.97 ± 0.44 | 3.72 ± 0.70 | 8.63 ± 0.83 |
| | $R^2_\star = 0.2$ | 66.62 ± 36.42 | 2.74 ± 0.65 | 4.69 ± 1.63 | 6.39 ± 3.18 | 8.25 ± 0.83 |
| | $R^2_\star = 0.4$ | 140.11 ± 42.70 | 3.95 ± 0.88 | 7.16 ± 1.22 | 10.90 ± 3.27 | 8.38 ± 1.09 |
| | $R^2_\star = 0.8$ | 1102.24 ± 400.08 | 8.82 ± 1.11 | 30.78 ± 5.14 | 74.20 ± 19.77 | 3.58 ± 1.28 |

Table 13: FinPILOT performance with controlled forecasts on DJ30, test year 2025, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | 15.65 ± 0.28 | 1.01 ± 0.02 | 1.34 ± 0.03 | 1.04 ± 0.02 | 15.23 ± 0.15 |
| | Original forecast | 23.68 ± 2.23 | 1.17 ± 0.09 | 1.61 ± 0.13 | 1.61 ± 0.11 | 14.42 ± 0.38 |
| | $R^2_\star = 0.005$ | 15.38 ± 1.35 | 0.85 ± 0.06 | 1.15 ± 0.09 | 1.11 ± 0.08 | 14.46 ± 0.39 |
| | $R^2_\star = 0.01$ | 17.11 ± 1.57 | 0.94 ± 0.07 | 1.26 ± 0.10 | 1.24 ± 0.09 | 14.24 ± 0.41 |
| | $R^2_\star = 0.025$ | 22.69 ± 1.56 | 1.20 ± 0.07 | 1.63 ± 0.09 | 1.64 ± 0.08 | 13.56 ± 0.46 |
| | $R^2_\star = 0.05$ | 33.12 ± 1.91 | 1.69 ± 0.07 | 2.31 ± 0.11 | 2.45 ± 0.12 | 12.36 ± 0.48 |
| | $R^2_\star = 0.1$ | 49.73 ± 1.65 | 2.44 ± 0.06 | 3.35 ± 0.18 | 3.85 ± 0.22 | 10.90 ± 0.44 |
| | $R^2_\star = 0.2$ | 96.02 ± 4.72 | 3.60 ± 0.15 | 5.15 ± 0.37 | 6.72 ± 0.61 | 10.34 ± 0.63 |
| | $R^2_\star = 0.4$ | 151.72 ± 1.68 | 3.97 ± 0.05 | 5.96 ± 0.12 | 8.05 ± 0.24 | 11.84 ± 0.30 |
| | $R^2_\star = 0.8$ | 282.87 ± 3.72 | 5.20 ± 0.01 | 8.90 ± 0.10 | 13.71 ± 0.34 | 10.07 ± 0.21 |
| **SAC** | Base | 16.09 ± 0.08 | 1.04 ± 0.00 | 1.37 ± 0.00 | 1.08 ± 0.00 | 14.92 ± 0.01 |
| | Original forecast | 17.73 ± 0.06 | 1.05 ± 0.00 | 1.44 ± 0.00 | 1.17 ± 0.00 | 15.15 ± 0.01 |
| | $R^2_\star = 0.005$ | 15.68 ± 0.05 | 0.95 ± 0.00 | 1.30 ± 0.00 | 1.05 ± 0.00 | 15.25 ± 0.01 |
| | $R^2_\star = 0.01$ | 16.08 ± 0.05 | 0.97 ± 0.00 | 1.33 ± 0.00 | 1.07 ± 0.00 | 15.22 ± 0.01 |
| | $R^2_\star = 0.025$ | 17.32 ± 0.05 | 1.03 ± 0.00 | 1.41 ± 0.00 | 1.15 ± 0.00 | 15.11 ± 0.01 |
| | $R^2_\star = 0.05$ | 19.69 ± 0.05 | 1.15 ± 0.00 | 1.57 ± 0.00 | 1.30 ± 0.00 | 14.98 ± 0.01 |
| | $R^2_\star = 0.1$ | 23.71 ± 0.07 | 1.35 ± 0.00 | 1.84 ± 0.00 | 1.56 ± 0.00 | 14.60 ± 0.01 |
| | $R^2_\star = 0.2$ | 30.41 ± 0.08 | 1.67 ± 0.00 | 2.28 ± 0.00 | 2.02 ± 0.01 | 13.82 ± 0.01 |
| | $R^2_\star = 0.4$ | 39.33 ± 0.08 | 2.08 ± 0.00 | 2.87 ± 0.00 | 2.75 ± 0.01 | 12.59 ± 0.02 |
| | $R^2_\star = 0.8$ | 46.28 ± 0.06 | 2.36 ± 0.00 | 3.32 ± 0.00 | 3.35 ± 0.01 | 11.78 ± 0.02 |
| **A2C** | Base | 16.11 ± 0.11 | 1.04 ± 0.00 | 1.37 ± 0.01 | 1.08 ± 0.01 | 14.96 ± 0.04 |
| | Original forecast | 25.11 ± 0.80 | 1.18 ± 0.03 | 1.76 ± 0.05 | 1.80 ± 0.06 | 13.67 ± 0.09 |
| | $R^2_\star = 0.005$ | 15.99 ± 0.64 | 0.84 ± 0.03 | 1.16 ± 0.04 | 1.15 ± 0.04 | 14.62 ± 0.13 |
| | $R^2_\star = 0.01$ | 17.62 ± 0.59 | 0.91 ± 0.03 | 1.26 ± 0.04 | 1.28 ± 0.05 | 14.27 ± 0.15 |
| | $R^2_\star = 0.025$ | 24.40 ± 0.48 | 1.21 ± 0.02 | 1.72 ± 0.04 | 1.82 ± 0.05 | 13.06 ± 0.17 |
| | $R^2_\star = 0.05$ | 36.56 ± 0.78 | 1.72 ± 0.03 | 2.46 ± 0.06 | 2.79 ± 0.10 | 11.87 ± 0.24 |
| | $R^2_\star = 0.1$ | 60.70 ± 0.59 | 2.64 ± 0.02 | 3.66 ± 0.03 | 4.53 ± 0.05 | 10.87 ± 0.05 |
| | $R^2_\star = 0.2$ | 100.71 ± 0.60 | 3.63 ± 0.01 | 5.41 ± 0.02 | 6.67 ± 0.04 | 10.75 ± 0.05 |
| | $R^2_\star = 0.4$ | 181.77 ± 0.73 | 4.57 ± 0.01 | 7.39 ± 0.04 | 10.69 ± 0.05 | 9.96 ± 0.03 |
| | $R^2_\star = 0.8$ | 340.94 ± 0.75 | 6.58 ± 0.01 | 12.70 ± 0.01 | 27.88 ± 0.05 | 5.43 ± 0.01 |
| **TD3** | Base | 15.82 ± 4.09 | 0.95 ± 0.15 | 1.28 ± 0.19 | 0.93 ± 0.13 | 17.09 ± 2.22 |
| | Original forecast | 28.56 ± 12.28 | 1.18 ± 0.40 | 1.80 ± 0.70 | 1.59 ± 0.61 | 18.38 ± 5.07 |
| | $R^2_\star = 0.005$ | 27.34 ± 9.44 | 1.03 ± 0.19 | 1.60 ± 0.40 | 1.48 ± 0.38 | 18.79 ± 3.54 |
| | $R^2_\star = 0.01$ | 23.05 ± 12.33 | 0.94 ± 0.37 | 1.43 ± 0.67 | 1.32 ± 0.74 | 19.93 ± 5.66 |
| | $R^2_\star = 0.025$ | 21.31 ± 4.12 | 0.93 ± 0.23 | 1.39 ± 0.31 | 1.22 ± 0.38 | 19.35 ± 4.10 |
| | $R^2_\star = 0.05$ | 40.38 ± 23.19 | 1.41 ± 0.45 | 2.15 ± 0.75 | 2.09 ± 1.13 | 17.56 ± 2.77 |
| | $R^2_\star = 0.1$ | 46.43 ± 34.17 | 1.53 ± 0.68 | 2.40 ± 1.32 | 2.39 ± 1.43 | 16.99 ± 4.04 |
| | $R^2_\star = 0.2$ | 78.92 ± 38.57 | 2.40 ± 0.65 | 4.01 ± 1.51 | 4.40 ± 2.19 | 14.29 ± 3.08 |
| | $R^2_\star = 0.4$ | 73.19 ± 52.79 | 2.50 ± 0.63 | 4.35 ± 1.95 | 5.89 ± 3.44 | 9.30 ± 1.58 |
| | $R^2_\star = 0.8$ | 1393.49 ± 576.38 | 6.49 ± 1.44 | 19.24 ± 5.86 | 34.29 ± 14.39 | 9.09 ± 3.25 |
| **DDPG** | Base | 12.94 ± 4.77 | 0.85 ± 0.29 | 1.10 ± 0.34 | 0.81 ± 0.26 | 16.73 ± 1.96 |
| | Original forecast | 20.38 ± 9.89 | 0.81 ± 0.32 | 1.29 ± 0.55 | 1.30 ± 0.57 | 17.42 ± 2.61 |
| | $R^2_\star = 0.005$ | 14.52 ± 10.45 | 0.68 ± 0.35 | 1.03 ± 0.56 | 0.80 ± 0.41 | 22.21 ± 4.92 |
| | $R^2_\star = 0.01$ | 17.40 ± 17.50 | 0.75 ± 0.63 | 1.13 ± 1.03 | 0.93 ± 0.88 | 24.13 ± 5.49 |
| | $R^2_\star = 0.025$ | 13.17 ± 7.18 | 0.63 ± 0.27 | 0.94 ± 0.42 | 0.83 ± 0.37 | 20.09 ± 4.94 |
| | $R^2_\star = 0.05$ | 35.93 ± 8.78 | 1.40 ± 0.25 | 2.08 ± 0.43 | 2.32 ± 0.43 | 14.54 ± 1.96 |
| | $R^2_\star = 0.1$ | 65.89 ± 6.70 | 1.90 ± 0.23 | 3.07 ± 0.49 | 3.71 ± 0.70 | 15.15 ± 2.77 |
| | $R^2_\star = 0.2$ | 196.86 ± 149.24 | 3.66 ± 0.87 | 7.51 ± 3.68 | 9.86 ± 6.53 | 11.64 ± 2.17 |
| | $R^2_\star = 0.4$ | 462.97 ± 355.25 | 4.18 ± 0.67 | 8.80 ± 2.42 | 13.54 ± 6.71 | 13.28 ± 4.93 |
| | $R^2_\star = 0.8$ | 2360.66 ± 1928.59 | 7.75 ± 1.55 | 25.06 ± 7.14 | 48.98 ± 19.01 | 6.64 ± 1.70 |

Table 14: FinPILOT performance with controlled forecasts on FX, test year 2023, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | 2.05 ± 0.13 | 0.39 ± 0.02 | 0.65 ± 0.04 | 0.40 ± 0.03 | 5.51 ± 0.22 |
| | Original forecast | 2.33 ± 0.97 | 0.54 ± 0.21 | 0.92 ± 0.33 | 0.45 ± 0.20 | 5.42 ± 0.52 |
| | $R^2_\star = 0.005$ | 1.44 ± 1.19 | 0.34 ± 0.26 | 0.58 ± 0.44 | 0.28 ± 0.23 | 5.98 ± 0.96 |
| | $R^2_\star = 0.01$ | 1.54 ± 1.20 | 0.36 ± 0.27 | 0.62 ± 0.44 | 0.30 ± 0.24 | 5.91 ± 0.97 |
| | $R^2_\star = 0.025$ | 1.86 ± 1.22 | 0.44 ± 0.27 | 0.74 ± 0.45 | 0.36 ± 0.25 | 5.76 ± 0.90 |
| | $R^2_\star = 0.05$ | 2.42 ± 1.14 | 0.56 ± 0.25 | 0.96 ± 0.40 | 0.47 ± 0.24 | 5.46 ± 0.63 |
| | $R^2_\star = 0.1$ | 3.75 ± 0.83 | 0.83 ± 0.21 | 1.39 ± 0.29 | 0.77 ± 0.19 | 4.93 ± 0.37 |
| | $R^2_\star = 0.2$ | 6.03 ± 0.69 | 1.23 ± 0.18 | 2.02 ± 0.24 | 1.17 ± 0.18 | 5.13 ± 0.40 |
| | $R^2_\star = 0.4$ | 13.32 ± 1.48 | 2.46 ± 0.28 | 4.05 ± 0.52 | 2.71 ± 0.33 | 4.67 ± 0.28 |
| | $R^2_\star = 0.8$ | 22.70 ± 1.44 | 3.28 ± 0.12 | 5.78 ± 0.37 | 3.89 ± 0.17 | 5.32 ± 0.22 |
| **SAC** | Base | 2.32 ± 0.02 | 0.45 ± 0.00 | 0.74 ± 0.01 | 0.46 ± 0.00 | 5.35 ± 0.01 |
| | Original forecast | 2.33 ± 0.02 | 0.49 ± 0.00 | 0.82 ± 0.01 | 0.46 ± 0.00 | 5.28 ± 0.01 |
| | $R^2_\star = 0.005$ | 2.13 ± 0.03 | 0.45 ± 0.00 | 0.75 ± 0.01 | 0.42 ± 0.00 | 5.33 ± 0.01 |
| | $R^2_\star = 0.01$ | 2.16 ± 0.03 | 0.46 ± 0.00 | 0.76 ± 0.01 | 0.42 ± 0.00 | 5.33 ± 0.01 |
| | $R^2_\star = 0.025$ | 2.25 ± 0.02 | 0.47 ± 0.00 | 0.79 ± 0.01 | 0.44 ± 0.00 | 5.33 ± 0.01 |
| | $R^2_\star = 0.05$ | 2.40 ± 0.02 | 0.50 ± 0.00 | 0.84 ± 0.01 | 0.47 ± 0.00 | 5.32 ± 0.01 |
| | $R^2_\star = 0.1$ | 2.70 ± 0.02 | 0.56 ± 0.00 | 0.94 ± 0.01 | 0.53 ± 0.00 | 5.30 ± 0.00 |
| | $R^2_\star = 0.2$ | 3.38 ± 0.02 | 0.69 ± 0.00 | 1.16 ± 0.01 | 0.66 ± 0.00 | 5.21 ± 0.01 |
| | $R^2_\star = 0.4$ | 4.88 ± 0.04 | 0.98 ± 0.01 | 1.65 ± 0.02 | 0.99 ± 0.01 | 4.96 ± 0.02 |
| | $R^2_\star = 0.8$ | 6.43 ± 0.06 | 1.21 ± 0.01 | 2.05 ± 0.02 | 1.29 ± 0.01 | 4.94 ± 0.02 |
| **A2C** | Base | 2.31 ± 0.03 | 0.45 ± 0.01 | 0.74 ± 0.01 | 0.46 ± 0.01 | 5.35 ± 0.01 |
| | Original forecast | 2.78 ± 0.03 | 0.61 ± 0.01 | 1.06 ± 0.02 | 0.54 ± 0.01 | 5.27 ± 0.05 |
| | $R^2_\star = 0.005$ | 1.87 ± 0.05 | 0.43 ± 0.01 | 0.74 ± 0.02 | 0.33 ± 0.01 | 6.03 ± 0.03 |
| | $R^2_\star = 0.01$ | 1.98 ± 0.03 | 0.45 ± 0.01 | 0.77 ± 0.01 | 0.35 ± 0.01 | 5.94 ± 0.03 |
| | $R^2_\star = 0.025$ | 2.27 ± 0.05 | 0.50 ± 0.02 | 0.85 ± 0.03 | 0.41 ± 0.01 | 5.77 ± 0.10 |
| | $R^2_\star = 0.05$ | 3.41 ± 0.06 | 0.71 ± 0.01 | 1.17 ± 0.02 | 0.61 ± 0.01 | 5.74 ± 0.07 |
| | $R^2_\star = 0.1$ | 4.70 ± 0.04 | 0.95 ± 0.01 | 1.56 ± 0.01 | 0.84 ± 0.01 | 5.64 ± 0.06 |
| | $R^2_\star = 0.2$ | 6.97 ± 0.08 | 1.36 ± 0.01 | 2.26 ± 0.01 | 1.27 ± 0.02 | 5.42 ± 0.05 |
| | $R^2_\star = 0.4$ | 15.06 ± 0.15 | 2.63 ± 0.04 | 4.34 ± 0.06 | 2.89 ± 0.06 | 4.91 ± 0.10 |
| | $R^2_\star = 0.8$ | 24.61 ± 0.13 | 3.18 ± 0.01 | 5.62 ± 0.03 | 3.42 ± 0.03 | 6.51 ± 0.05 |
| **TD3** | Base | 2.63 ± 0.84 | 0.51 ± 0.15 | 0.82 ± 0.23 | 0.50 ± 0.19 | 5.67 ± 0.63 |
| | Original forecast | 4.76 ± 1.78 | 0.90 ± 0.27 | 1.49 ± 0.47 | 0.75 ± 0.26 | 6.45 ± 1.20 |
| | $R^2_\star = 0.005$ | 2.28 ± 2.10 | 0.41 ± 0.36 | 0.65 ± 0.60 | 0.33 ± 0.31 | 7.71 ± 1.28 |
| | $R^2_\star = 0.01$ | 5.43 ± 2.35 | 1.00 ± 0.46 | 1.56 ± 0.73 | 0.85 ± 0.49 | 6.90 ± 1.09 |
| | $R^2_\star = 0.025$ | 5.28 ± 4.02 | 0.92 ± 0.75 | 1.51 ± 1.17 | 0.78 ± 0.55 | 7.35 ± 1.21 |
| | $R^2_\star = 0.05$ | 6.48 ± 1.95 | 1.14 ± 0.32 | 1.79 ± 0.49 | 0.90 ± 0.25 | 7.17 ± 0.26 |
| | $R^2_\star = 0.1$ | 9.33 ± 5.16 | 1.53 ± 0.75 | 2.45 ± 1.36 | 1.35 ± 0.70 | 6.42 ± 0.99 |
| | $R^2_\star = 0.2$ | 12.19 ± 1.62 | 1.86 ± 0.28 | 2.91 ± 0.63 | 1.63 ± 0.22 | 7.17 ± 0.32 |
| | $R^2_\star = 0.4$ | 28.96 ± 9.42 | 3.05 ± 0.45 | 4.78 ± 1.08 | 4.31 ± 1.47 | 6.10 ± 1.02 |
| | $R^2_\star = 0.8$ | 66.77 ± 7.57 | 4.81 ± 0.65 | 9.90 ± 2.07 | 11.87 ± 4.09 | 4.73 ± 1.34 |
| **DDPG** | Base | 2.62 ± 1.77 | 0.51 ± 0.31 | 0.81 ± 0.51 | 0.50 ± 0.30 | 5.70 ± 0.83 |
| | Original forecast | 4.96 ± 2.60 | 1.00 ± 0.44 | 1.63 ± 0.80 | 0.79 ± 0.38 | 6.31 ± 0.83 |
| | $R^2_\star = 0.005$ | 3.75 ± 2.66 | 0.73 ± 0.42 | 1.18 ± 0.71 | 0.55 ± 0.35 | 6.69 ± 0.45 |
| | $R^2_\star = 0.01$ | 4.45 ± 3.90 | 0.83 ± 0.54 | 1.38 ± 1.00 | 0.66 ± 0.52 | 6.54 ± 0.48 |
| | $R^2_\star = 0.025$ | 6.12 ± 3.74 | 1.10 ± 0.49 | 1.74 ± 0.81 | 0.90 ± 0.45 | 6.41 ± 0.84 |
| | $R^2_\star = 0.05$ | 7.29 ± 4.62 | 1.27 ± 0.63 | 2.01 ± 1.11 | 1.01 ± 0.55 | 6.80 ± 0.69 |
| | $R^2_\star = 0.1$ | 8.65 ± 1.71 | 1.50 ± 0.29 | 2.32 ± 0.56 | 1.28 ± 0.26 | 6.71 ± 1.17 |
| | $R^2_\star = 0.2$ | 12.75 ± 2.69 | 1.96 ± 0.22 | 3.12 ± 0.42 | 1.92 ± 0.44 | 6.45 ± 1.04 |
| | $R^2_\star = 0.4$ | 27.15 ± 3.87 | 3.00 ± 0.37 | 4.67 ± 0.80 | 3.97 ± 1.17 | 6.45 ± 1.44 |
| | $R^2_\star = 0.8$ | 54.27 ± 12.83 | 4.14 ± 0.57 | 8.26 ± 1.79 | 8.20 ± 2.72 | 5.60 ± 1.19 |

Table 15: FinPILOT performance with controlled forecasts on FX, test year 2024, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | -5.82 ± 0.10 | -1.36 ± 0.01 | -1.87 ± 0.02 | -0.90 ± 0.01 | 6.53 ± 0.10 |
| | Original forecast | -4.60 ± 0.24 | -1.24 ± 0.05 | -1.67 ± 0.08 | -0.95 ± 0.04 | 4.85 ± 0.34 |
| | $R^2_\star = 0.005$ | -4.67 ± 0.20 | -1.25 ± 0.07 | -1.69 ± 0.10 | -0.93 ± 0.05 | 4.99 ± 0.44 |
| | $R^2_\star = 0.01$ | -4.61 ± 0.19 | -1.23 ± 0.07 | -1.67 ± 0.09 | -0.92 ± 0.06 | 4.99 ± 0.47 |
| | $R^2_\star = 0.025$ | -4.46 ± 0.10 | -1.18 ± 0.05 | -1.61 ± 0.07 | -0.89 ± 0.07 | 5.05 ± 0.47 |
| | $R^2_\star = 0.05$ | -4.16 ± 0.14 | -1.10 ± 0.05 | -1.51 ± 0.06 | -0.83 ± 0.08 | 5.05 ± 0.47 |
| | $R^2_\star = 0.1$ | -3.75 ± 0.23 | -0.98 ± 0.09 | -1.36 ± 0.12 | -0.73 ± 0.07 | 5.16 ± 0.29 |
| | $R^2_\star = 0.2$ | -2.77 ± 0.21 | -0.72 ± 0.04 | -1.01 ± 0.05 | -0.53 ± 0.04 | 5.20 ± 0.30 |
| | $R^2_\star = 0.4$ | -0.93 ± 0.42 | -0.24 ± 0.11 | -0.34 ± 0.15 | -0.18 ± 0.08 | 4.91 ± 0.56 |
| | $R^2_\star = 0.8$ | 7.83 ± 1.15 | 1.74 ± 0.25 | 2.62 ± 0.38 | 1.76 ± 0.35 | 4.40 ± 0.32 |
| **SAC** | Base | -5.74 ± 0.07 | -1.35 ± 0.01 | -1.87 ± 0.02 | -0.90 ± 0.01 | 6.49 ± 0.04 |
| | Original forecast | -5.53 ± 0.05 | -1.42 ± 0.01 | -1.89 ± 0.01 | -1.00 ± 0.01 | 5.60 ± 0.01 |
| | $R^2_\star = 0.005$ | -5.46 ± 0.05 | -1.41 ± 0.01 | -1.86 ± 0.01 | -0.98 ± 0.01 | 5.64 ± 0.01 |
| | $R^2_\star = 0.01$ | -5.40 ± 0.05 | -1.39 ± 0.01 | -1.84 ± 0.01 | -0.97 ± 0.01 | 5.64 ± 0.02 |
| | $R^2_\star = 0.025$ | -5.24 ± 0.05 | -1.35 ± 0.01 | -1.78 ± 0.01 | -0.94 ± 0.01 | 5.63 ± 0.01 |
| | $R^2_\star = 0.05$ | -4.97 ± 0.05 | -1.28 ± 0.01 | -1.69 ± 0.01 | -0.90 ± 0.01 | 5.60 ± 0.01 |
| | $R^2_\star = 0.1$ | -4.46 ± 0.05 | -1.14 ± 0.01 | -1.51 ± 0.01 | -0.81 ± 0.01 | 5.57 ± 0.01 |
| | $R^2_\star = 0.2$ | -3.56 ± 0.04 | -0.90 ± 0.01 | -1.20 ± 0.01 | -0.65 ± 0.01 | 5.48 ± 0.01 |
| | $R^2_\star = 0.4$ | -2.25 ± 0.03 | -0.55 ± 0.01 | -0.75 ± 0.01 | -0.42 ± 0.01 | 5.28 ± 0.01 |
| | $R^2_\star = 0.8$ | -0.88 ± 0.03 | -0.20 ± 0.01 | -0.27 ± 0.01 | -0.16 ± 0.00 | 5.12 ± 0.01 |
| **A2C** | Base | -5.78 ± 0.00 | -1.36 ± 0.00 | -1.88 ± 0.00 | -0.90 ± 0.00 | 6.51 ± 0.01 |
| | Original forecast | -3.79 ± 0.24 | -1.05 ± 0.05 | -1.43 ± 0.07 | -0.84 ± 0.02 | 4.51 ± 0.19 |
| | $R^2_\star = 0.005$ | -4.37 ± 0.10 | -1.19 ± 0.02 | -1.60 ± 0.03 | -0.89 ± 0.01 | 4.93 ± 0.08 |
| | $R^2_\star = 0.01$ | -4.36 ± 0.13 | -1.18 ± 0.03 | -1.59 ± 0.04 | -0.88 ± 0.01 | 4.99 ± 0.15 |
| | $R^2_\star = 0.025$ | -4.35 ± 0.16 | -1.16 ± 0.04 | -1.59 ± 0.06 | -0.85 ± 0.01 | 5.17 ± 0.15 |
| | $R^2_\star = 0.05$ | -3.91 ± 0.11 | -1.03 ± 0.03 | -1.44 ± 0.04 | -0.75 ± 0.02 | 5.20 ± 0.04 |
| | $R^2_\star = 0.1$ | -3.39 ± 0.10 | -0.89 ± 0.03 | -1.24 ± 0.04 | -0.65 ± 0.02 | 5.20 ± 0.04 |
| | $R^2_\star = 0.2$ | -2.53 ± 0.08 | -0.67 ± 0.01 | -0.94 ± 0.01 | -0.48 ± 0.01 | 5.17 ± 0.09 |
| | $R^2_\star = 0.4$ | -0.44 ± 0.20 | -0.12 ± 0.06 | -0.18 ± 0.09 | -0.09 ± 0.05 | 4.12 ± 0.15 |
| | $R^2_\star = 0.8$ | 7.38 ± 0.24 | 1.69 ± 0.05 | 2.56 ± 0.08 | 1.68 ± 0.06 | 4.28 ± 0.03 |
| **TD3** | Base | -7.36 ± 1.37 | -1.56 ± 0.24 | -2.24 ± 0.27 | -0.99 ± 0.05 | 7.43 ± 1.21 |
| | Original forecast | -3.23 ± 1.51 | -0.83 ± 0.33 | -1.18 ± 0.39 | -0.67 ± 0.14 | 4.55 ± 1.44 |
| | $R^2_\star = 0.005$ | -4.84 ± 3.54 | -1.08 ± 0.52 | -1.47 ± 0.59 | -0.79 ± 0.15 | 5.64 ± 3.15 |
| | $R^2_\star = 0.01$ | -3.92 ± 2.46 | -0.93 ± 0.39 | -1.27 ± 0.44 | -0.73 ± 0.15 | 4.93 ± 2.10 |
| | $R^2_\star = 0.025$ | -4.48 ± 3.66 | -1.00 ± 0.55 | -1.36 ± 0.66 | -0.74 ± 0.19 | 5.43 ± 3.08 |
| | $R^2_\star = 0.05$ | -2.81 ± 0.97 | -0.76 ± 0.22 | -1.09 ± 0.25 | -0.66 ± 0.14 | 4.08 ± 0.71 |
| | $R^2_\star = 0.1$ | -0.73 ± 0.98 | -0.22 ± 0.29 | -0.32 ± 0.43 | -0.18 ± 0.30 | 3.34 ± 0.71 |
| | $R^2_\star = 0.2$ | 3.03 ± 2.46 | 0.66 ± 0.30 | 0.90 ± 0.36 | 0.66 ± 0.18 | 4.41 ± 3.02 |
| | $R^2_\star = 0.4$ | 12.90 ± 4.91 | 2.17 ± 1.06 | 3.33 ± 1.91 | 4.03 ± 3.79 | 5.09 ± 2.97 |
| | $R^2_\star = 0.8$ | 17.26 ± 7.15 | 2.47 ± 0.95 | 3.78 ± 1.76 | 4.86 ± 3.38 | 4.61 ± 2.73 |
| **DDPG** | Base | -6.78 ± 1.31 | -1.62 ± 0.28 | -2.33 ± 0.49 | -1.01 ± 0.03 | 6.68 ± 1.20 |
| | Original forecast | -4.36 ± 1.58 | -1.25 ± 0.48 | -1.61 ± 0.48 | -0.87 ± 0.15 | 4.93 ± 1.48 |
| | $R^2_\star = 0.005$ | -4.61 ± 3.12 | -1.22 ± 0.73 | -1.51 ± 0.65 | -0.78 ± 0.22 | 5.58 ± 2.66 |
| | $R^2_\star = 0.01$ | -3.83 ± 2.18 | -1.09 ± 0.67 | -1.44 ± 0.73 | -0.74 ± 0.26 | 4.78 ± 1.40 |
| | $R^2_\star = 0.025$ | -3.32 ± 1.38 | -0.90 ± 0.30 | -1.25 ± 0.32 | -0.72 ± 0.17 | 4.51 ± 1.27 |
| | $R^2_\star = 0.05$ | -3.04 ± 1.34 | -0.76 ± 0.27 | -1.10 ± 0.31 | -0.64 ± 0.07 | 4.60 ± 1.54 |
| | $R^2_\star = 0.1$ | -0.01 ± 1.49 | 0.03 ± 0.44 | 0.06 ± 0.74 | 0.14 ± 0.67 | 3.18 ± 0.78 |
| | $R^2_\star = 0.2$ | 4.83 ± 2.24 | 1.21 ± 0.89 | 1.80 ± 1.49 | 1.49 ± 1.90 | 5.39 ± 2.61 |
| | $R^2_\star = 0.4$ | 15.53 ± 7.56 | 2.45 ± 1.22 | 3.54 ± 1.87 | 4.24 ± 2.98 | 5.03 ± 2.88 |
| | $R^2_\star = 0.8$ | 22.37 ± 7.44 | 3.69 ± 0.60 | 5.70 ± 0.97 | 9.17 ± 3.72 | 2.59 ± 1.30 |

Table 16: FinPILOT performance with controlled forecasts on FX, test year 2025, at target skill levels from $R^2_\star = 0.005$ to $R^2_\star = 0.8$ (Section 4.4). Values are mean $\pm$ one standard deviation across 5 seeds. Base is the frozen pre-trained policy and does not depend on $R^2_\star$; Original forecast is the unmodified forecast ($c_{t,i} = 1$).

| Algorithm | $R^2_\star$ | Total return (%) ↑ | Sharpe ↑ | Sortino ↑ | Calmar ↑ | MDD (%) ↓ |
|---|---|---|---|---|---|---|
| **PPO** | Base | $7.00 \pm 0.26$ | $1.40 \pm 0.04$ | $2.41 \pm 0.08$ | $2.70 \pm 0.11$ | $2.55 \pm 0.04$ |
| | Original forecast | $6.14 \pm 0.78$ | $1.54 \pm 0.15$ | $3.05 \pm 0.29$ | $3.31 \pm 0.85$ | $1.91 \pm 0.45$ |
| | $R^2_\star = 0.005$ | $6.35 \pm 0.80$ | $1.54 \pm 0.16$ | $3.02 \pm 0.32$ | $3.27 \pm 0.86$ | $1.99 \pm 0.43$ |
| | $R^2_\star = 0.01$ | $6.53 \pm 0.79$ | $1.56 \pm 0.15$ | $3.08 \pm 0.30$ | $3.28 \pm 0.83$ | $2.03 \pm 0.41$ |
| | $R^2_\star = 0.025$ | $7.21 \pm 0.75$ | $1.65 \pm 0.13$ | $3.30 \pm 0.27$ | $3.55 \pm 0.86$ | $2.06 \pm 0.41$ |
| | $R^2_\star = 0.05$ | $8.28 \pm 0.70$ | $1.78 \pm 0.14$ | $3.63 \pm 0.30$ | $4.03 \pm 0.96$ | $2.07 \pm 0.40$ |
| | $R^2_\star = 0.1$ | $10.20 \pm 0.66$ | $1.94 \pm 0.12$ | $4.04 \pm 0.30$ | $4.44 \pm 0.72$ | $2.25 \pm 0.27$ |
| | $R^2_\star = 0.2$ | $13.83 \pm 0.52$ | $2.27 \pm 0.07$ | $4.90 \pm 0.27$ | $5.63 \pm 0.58$ | $2.35 \pm 0.20$ |
| | $R^2_\star = 0.4$ | $18.82 \pm 1.01$ | $2.64 \pm 0.14$ | $5.59 \pm 0.46$ | $6.91 \pm 0.86$ | $2.56 \pm 0.29$ |
| | $R^2_\star = 0.8$ | $21.25 \pm 0.47$ | $2.71 \pm 0.05$ | $5.43 \pm 0.15$ | $6.10 \pm 0.20$ | $3.20 \pm 0.06$ |
| **SAC** | Base | $7.02 \pm 0.07$ | $1.39 \pm 0.00$ | $2.39 \pm 0.01$ | $2.70 \pm 0.01$ | $2.56 \pm 0.03$ |
| | Original forecast | $7.06 \pm 0.04$ | $1.54 \pm 0.01$ | $2.80 \pm 0.02$ | $3.07 \pm 0.01$ | $2.25 \pm 0.00$ |
| | $R^2_\star = 0.005$ | $7.00 \pm 0.05$ | $1.52 \pm 0.01$ | $2.75 \pm 0.03$ | $3.01 \pm 0.01$ | $2.28 \pm 0.00$ |
| | $R^2_\star = 0.01$ | $7.05 \pm 0.05$ | $1.53 \pm 0.01$ | $2.77 \pm 0.03$ | $3.03 \pm 0.02$ | $2.28 \pm 0.00$ |
| | $R^2_\star = 0.025$ | $7.20 \pm 0.04$ | $1.55 \pm 0.01$ | $2.82 \pm 0.02$ | $3.09 \pm 0.01$ | $2.28 \pm 0.00$ |
| | $R^2_\star = 0.05$ | $7.47 \pm 0.05$ | $1.58 \pm 0.01$ | $2.90 \pm 0.03$ | $3.20 \pm 0.02$ | $2.28 \pm 0.00$ |
| | $R^2_\star = 0.1$ | $8.00 \pm 0.06$ | $1.65 \pm 0.01$ | $3.03 \pm 0.03$ | $3.40 \pm 0.02$ | $2.29 \pm 0.00$ |
| | $R^2_\star = 0.2$ | $9.11 \pm 0.07$ | $1.76 \pm 0.01$ | $3.24 \pm 0.02$ | $3.79 \pm 0.02$ | $2.34 \pm 0.01$ |
| | $R^2_\star = 0.4$ | $10.79 \pm 0.07$ | $1.93 \pm 0.01$ | $3.54 \pm 0.02$ | $4.51 \pm 0.02$ | $2.31 \pm 0.00$ |
| | $R^2_\star = 0.8$ | $12.56 \pm 0.07$ | $2.14 \pm 0.01$ | $3.94 \pm 0.02$ | $5.39 \pm 0.02$ | $2.22 \pm 0.00$ |
| **A2C** | Base | $7.06 \pm 0.02$ | $1.39 \pm 0.00$ | $2.39 \pm 0.01$ | $2.71 \pm 0.01$ | $2.57 \pm 0.01$ |
| | Original forecast | $5.34 \pm 0.08$ | $1.43 \pm 0.01$ | $2.93 \pm 0.02$ | $2.66 \pm 0.04$ | $1.98 \pm 0.01$ |
| | $R^2_\star = 0.005$ | $5.67 \pm 0.09$ | $1.43 \pm 0.01$ | $2.95 \pm 0.03$ | $2.67 \pm 0.04$ | $2.10 \pm 0.00$ |
| | $R^2_\star = 0.01$ | $5.92 \pm 0.11$ | $1.47 \pm 0.02$ | $3.05 \pm 0.04$ | $2.77 \pm 0.05$ | $2.11 \pm 0.00$ |
| | $R^2_\star = 0.025$ | $6.63 \pm 0.08$ | $1.58 \pm 0.01$ | $3.29 \pm 0.01$ | $3.06 \pm 0.04$ | $2.13 \pm 0.01$ |
| | $R^2_\star = 0.05$ | $7.58 \pm 0.07$ | $1.69 \pm 0.01$ | $3.50 \pm 0.04$ | $3.33 \pm 0.04$ | $2.22 \pm 0.03$ |
| | $R^2_\star = 0.1$ | $9.54 \pm 0.05$ | $1.91 \pm 0.01$ | $3.92 \pm 0.01$ | $4.46 \pm 0.07$ | $2.07 \pm 0.03$ |
| | $R^2_\star = 0.2$ | $14.77 \pm 0.32$ | $2.36 \pm 0.03$ | $5.19 \pm 0.06$ | $6.41 \pm 0.16$ | $2.18 \pm 0.07$ |
| | $R^2_\star = 0.4$ | $23.23 \pm 0.09$ | $3.09 \pm 0.01$ | $7.52 \pm 0.03$ | $9.05 \pm 0.08$ | $2.33 \pm 0.02$ |
| | $R^2_\star = 0.8$ | $26.68 \pm 0.04$ | $3.19 \pm 0.01$ | $7.16 \pm 0.02$ | $7.69 \pm 0.07$ | $3.11 \pm 0.03$ |
| **TD3** | Base | $6.35 \pm 1.05$ | $1.33 \pm 0.19$ | $2.32 \pm 0.29$ | $2.64 \pm 0.74$ | $2.47 \pm 0.47$ |
| | Original forecast | $8.74 \pm 0.86$ | $1.53 \pm 0.09$ | $2.86 \pm 0.18$ | $3.55 \pm 0.18$ | $2.40 \pm 0.22$ |
| | $R^2_\star = 0.005$ | $6.63 \pm 2.13$ | $1.26 \pm 0.30$ | $2.35 \pm 0.56$ | $2.46 \pm 1.06$ | $2.87 \pm 0.70$ |
| | $R^2_\star = 0.01$ | $6.81 \pm 2.35$ | $1.27 \pm 0.35$ | $2.39 \pm 0.65$ | $2.45 \pm 1.14$ | $2.94 \pm 0.60$ |
| | $R^2_\star = 0.025$ | $6.37 \pm 1.72$ | $1.19 \pm 0.25$ | $2.34 \pm 0.55$ | $2.25 \pm 0.80$ | $2.92 \pm 0.45$ |
| | $R^2_\star = 0.05$ | $9.35 \pm 2.24$ | $1.56 \pm 0.26$ | $3.13 \pm 0.54$ | $3.66 \pm 1.15$ | $2.56 \pm 0.37$ |
| | $R^2_\star = 0.1$ | $11.26 \pm 0.96$ | $1.79 \pm 0.21$ | $3.54 \pm 0.42$ | $4.52 \pm 0.99$ | $2.46 \pm 0.34$ |
| | $R^2_\star = 0.2$ | $15.73 \pm 2.72$ | $2.14 \pm 0.19$ | $4.56 \pm 0.57$ | $5.77 \pm 1.36$ | $2.62 \pm 0.35$ |
| | $R^2_\star = 0.4$ | $24.97 \pm 6.59$ | $2.70 \pm 0.44$ | $5.61 \pm 1.70$ | $8.09 \pm 2.43$ | $2.87 \pm 0.47$ |
| | $R^2_\star = 0.8$ | $35.56 \pm 11.51$ | $3.41 \pm 0.89$ | $7.49 \pm 3.59$ | $10.58 \pm 4.76$ | $3.11 \pm 0.71$ |
| **DDPG** | Base | $3.52 \pm 1.85$ | $0.79 \pm 0.39$ | $1.35 \pm 0.67$ | $1.33 \pm 0.97$ | $3.18 \pm 1.10$ |
| | Original forecast | $3.84 \pm 0.91$ | $0.82 \pm 0.17$ | $1.51 \pm 0.28$ | $1.07 \pm 0.39$ | $3.76 \pm 0.62$ |
| | $R^2_\star = 0.005$ | $2.82 \pm 1.69$ | $0.65 \pm 0.39$ | $1.17 \pm 0.70$ | $0.91 \pm 0.67$ | $3.99 \pm 1.87$ |
| | $R^2_\star = 0.01$ | $2.53 \pm 1.10$ | $0.57 \pm 0.23$ | $1.05 \pm 0.43$ | $0.70 \pm 0.30$ | $3.78 \pm 0.74$ |
| | $R^2_\star = 0.025$ | $3.19 \pm 1.43$ | $0.69 \pm 0.28$ | $1.27 \pm 0.55$ | $0.96 \pm 0.55$ | $3.60 \pm 0.85$ |
| | $R^2_\star = 0.05$ | $4.61 \pm 1.41$ | $0.97 \pm 0.24$ | $1.84 \pm 0.49$ | $1.65 \pm 0.56$ | $2.87 \pm 0.53$ |
| | $R^2_\star = 0.1$ | $8.71 \pm 3.17$ | $1.54 \pm 0.51$ | $3.11 \pm 1.09$ | $3.19 \pm 1.79$ | $3.05 \pm 1.06$ |
| | $R^2_\star = 0.2$ | $14.90 \pm 5.38$ | $2.07 \pm 0.45$ | $4.49 \pm 1.04$ | $4.70 \pm 2.41$ | $3.27 \pm 1.03$ |
| | $R^2_\star = 0.4$ | $21.18 \pm 4.11$ | $2.53 \pm 0.41$ | $5.45 \pm 1.18$ | $6.60 \pm 2.33$ | $3.13 \pm 0.77$ |
| | $R^2_\star = 0.8$ | $30.65 \pm 7.20$ | $3.26 \pm 0.44$ | $6.51 \pm 1.01$ | $10.41 \pm 3.97$ | $2.79 \pm 0.74$ |

### D.4 SENSITIVITY TO STEP SIZE AND ADAPTATION EPOCHS

We vary the step size $\alpha \in \{0.001, 0.003, 0.01, 0.03, 0.1\}$ (with $E = 2$) and the number of adaptation epochs $E \in \{1, 2, 3, 4\}$ (with $\alpha = 0.01$), holding all other settings, including the selected $\lambda$ values and seeds, at those of the main experiments. Both analyses are post-hoc and were not used to select any setting. Table 17 reports averages over the 15 algorithm–years per market, and Figure 5 breaks ΔSharpe down by test year.

Table 17: Post-hoc sensitivity of FinPILOT to the step size $\alpha$ (with $E = 2$) and the number of adaptation epochs $E$ (with $\alpha = 0.01$). All other settings match the main experiments. Values are changes relative to Base, averaged over the 15 algorithm–years per market; the subscript is the seed-level standard deviation of the change, computed from paired per-seed differences (FinPILOT minus Base on the same seed) within each algorithm–year and pooled over the 15 algorithm–years. Sharpe↑ counts algorithm–years where Sharpe improves, and Best counts, within each sweep, the algorithm–years in which that setting gives the highest ΔSharpe among the values tested. The main experiments use $\alpha = 0.01$ and $E = 2$; per-year results are shown in Figure 5.

| Setting | ΔRet. (pp) | ΔSharpe | ΔSortino | ΔCalmar | ΔMDD (pp)↓ | Sharpe↑ | Best |
|---|---|---|---|---|---|---|---|
| *DJ30* | | | | | | | |
| $\alpha = 0.001$ | $+2.08_{\pm 2.80}$ | $+0.07_{\pm 0.16}$ | $+0.13_{\pm 0.25}$ | $+0.19_{\pm 0.24}$ | $+0.31_{\pm 0.85}$ | 12/15 | 0/15 |
| $\alpha = 0.003$ | $+8.89_{\pm 6.93}$ | $+0.22_{\pm 0.26}$ | $+0.39_{\pm 0.39}$ | $+0.50_{\pm 0.55}$ | $+1.43_{\pm 2.10}$ | 15/15 | 3/15 |
| $\alpha = 0.01$ | $+14.92_{\pm 6.15}$ | $+0.38_{\pm 0.30}$ | $+0.70_{\pm 0.51}$ | $+0.90_{\pm 0.66}$ | $+1.57_{\pm 2.42}$ | 13/15 | 7/15 |
| $\alpha = 0.03$ | $+14.20_{\pm 9.22}$ | $+0.24_{\pm 0.37}$ | $+0.54_{\pm 0.62}$ | $+0.71_{\pm 0.71}$ | $+3.26_{\pm 2.78}$ | 10/15 | 1/15 |
| $\alpha = 0.1$ | $+8.60_{\pm 5.41}$ | $+0.03_{\pm 0.30}$ | $+0.18_{\pm 0.47}$ | $+0.25_{\pm 0.54}$ | $+3.37_{\pm 2.28}$ | 7/15 | 4/15 |
| $E = 1$ | $+9.75_{\pm 7.01}$ | $+0.23_{\pm 0.29}$ | $+0.43_{\pm 0.41}$ | $+0.62_{\pm 0.65}$ | $+1.44_{\pm 2.63}$ | 14/15 | 3/15 |
| $E = 2$ | $+14.92_{\pm 6.15}$ | $+0.38_{\pm 0.30}$ | $+0.70_{\pm 0.51}$ | $+0.90_{\pm 0.66}$ | $+1.57_{\pm 2.42}$ | 13/15 | 6/15 |
| $E = 3$ | $+13.70_{\pm 11.13}$ | $+0.23_{\pm 0.51}$ | $+0.51_{\pm 0.73}$ | $+0.66_{\pm 1.06}$ | $+3.25_{\pm 3.68}$ | 10/15 | 2/15 |
| $E = 4$ | $+13.93_{\pm 12.76}$ | $+0.19_{\pm 0.49}$ | $+0.43_{\pm 0.74}$ | $+0.64_{\pm 1.04}$ | $+3.63_{\pm 2.85}$ | 9/15 | 4/15 |
| *FX* | | | | | | | |
| $\alpha = 0.001$ | $+0.40_{\pm 0.69}$ | $+0.07_{\pm 0.14}$ | $+0.16_{\pm 0.23}$ | $+0.07_{\pm 0.20}$ | $-0.38_{\pm 0.51}$ | 12/15 | 1/15 |
| $\alpha = 0.003$ | $+0.42_{\pm 1.23}$ | $+0.07_{\pm 0.25}$ | $+0.20_{\pm 0.40}$ | $-0.06_{\pm 0.56}$ | $-0.26_{\pm 0.64}$ | 9/15 | 0/15 |
| $\alpha = 0.01$ | $+1.02_{\pm 1.15}$ | $+0.22_{\pm 0.25}$ | $+0.46_{\pm 0.38}$ | $+0.18_{\pm 0.40}$ | $-0.61_{\pm 0.93}$ | 14/15 | 2/15 |
| $\alpha = 0.03$ | $+0.83_{\pm 1.35}$ | $+0.22_{\pm 0.29}$ | $+0.45_{\pm 0.48}$ | $+0.27_{\pm 0.45}$ | $-0.68_{\pm 0.95}$ | 13/15 | 2/15 |
| $\alpha = 0.1$ | $+1.55_{\pm 1.06}$ | $+0.41_{\pm 0.17}$ | $+0.75_{\pm 0.30}$ | $+0.70_{\pm 0.30}$ | $-0.97_{\pm 1.00}$ | 14/15 | 10/15 |
| $E = 1$ | $+0.51_{\pm 1.08}$ | $+0.10_{\pm 0.20}$ | $+0.24_{\pm 0.32}$ | $-0.16_{\pm 0.41}$ | $-0.32_{\pm 0.57}$ | 9/15 | 3/15 |
| $E = 2$ | $+1.02_{\pm 1.15}$ | $+0.22_{\pm 0.25}$ | $+0.46_{\pm 0.38}$ | $+0.18_{\pm 0.40}$ | $-0.61_{\pm 0.93}$ | 14/15 | 4/15 |
| $E = 3$ | $+1.00_{\pm 1.57}$ | $+0.21_{\pm 0.32}$ | $+0.44_{\pm 0.49}$ | $+0.17_{\pm 0.51}$ | $-0.74_{\pm 0.90}$ | 13/15 | 3/15 |
| $E = 4$ | $+0.64_{\pm 1.53}$ | $+0.17_{\pm 0.28}$ | $+0.38_{\pm 0.42}$ | $+0.11_{\pm 0.44}$ | $-0.45_{\pm 0.84}$ | 12/15 | 5/15 |

**Step size.** On DJ30, gains rise with the step size up to $\alpha = 0.01$ and flatten or decline beyond it: mean return is similar at $\alpha = 0.01$ and 0.03 ($+14.92$ vs. $+14.20$ pp), but Sharpe falls from $+0.38$ to $+0.24$ and the increase in maximum drawdown over Base roughly doubles ($+1.57$ to $+3.26$ pp). At $\alpha = 0.1$, mean ΔSharpe drops to $+0.03$. The decline is concentrated in 2024, the year with the weakest forecaster, where Sharpe falls below the base policy for all five algorithms at $\alpha \geq 0.03$. On FX, gains generally increase with $\alpha$: $\alpha = 0.1$ attains the largest mean improvement on every metric, raising mean return from $+1.02$ to $+1.55$ pp, so our shared step size is conservative for this market.

**Adaptation epochs.** With $E = 1$, FinPILOT retains most of its DJ30 gain ($+9.75$ vs. $+14.92$ pp) but roughly halves the FX gain, improving FX Sharpe in only 9 of 15 algorithm–years. In DJ30 2023, $E \in \{3, 4\}$ gives the largest returns for every algorithm, whereas in DJ30 2024 Sharpe falls below the base policy for four of five algorithms at $E = 3$ and all five at $E = 4$, and the mean DJ30 drawdown increase over Base roughly doubles relative to $E = 2$ ($+1.57$ to $+3.25$ and $+3.63$ pp). On FX, $E = 2$ and $E = 3$ perform similarly. Averaged over algorithm–years, $E = 2$ attains the highest mean return and Sharpe in both markets among the values tested, although it is best in only a minority of individual algorithm–years.

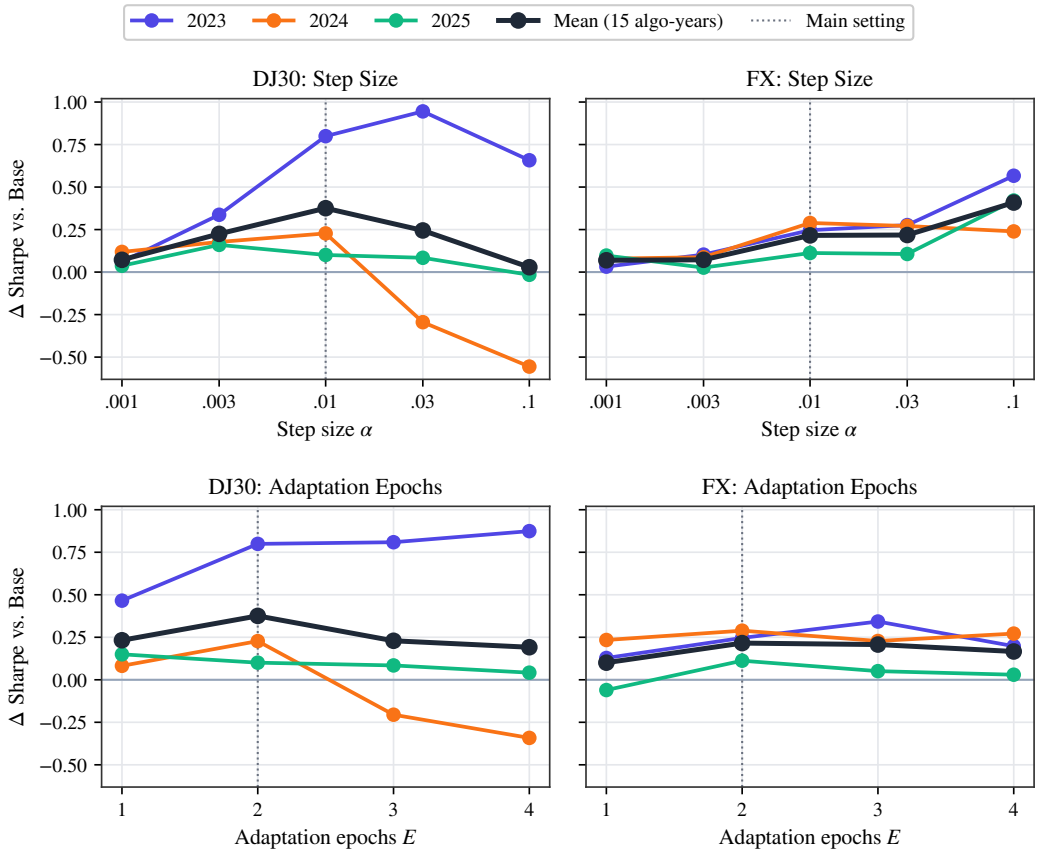



Figure 5: Post-hoc sensitivity of FinPILOT to the step size $\alpha$ (top; $E = 2$) and the number of adaptation epochs $E$ (bottom; $\alpha = 0.01$). Colored lines show the mean $\Delta$Sharpe over the five algorithms in each test year; the black line is the mean over all 15 algorithm–years in each market. The dotted line marks the setting used in all main results. Pooled forecaster $R^2$ for 2023/2024/2025 is 0.027/0.005/0.014 on DJ30 and 0.039/0.029/0.019 on FX (Table 6). Lines show means only; seed-level variability at the main setting is reported in Tables 7 and 8.

Mean return and Sharpe improvements over Base are positive at every tested value of $\alpha$ and $E$ in both markets: these settings affect the magnitude of the gains but not their direction. Both sweeps are consistent with the appropriate magnitude of adaptation depending on forecast skill: the clearest degradation from larger updates occurs in DJ30 2024, the year with the weakest forecaster ($R^2 = 0.005$).