Highlights

- We release ML Trading Bench, a one-command reproducible benchmark for ML-based quantitative trading on 50 US ETFs (2005–2024), producing 9 tables and 9 figures in under 6 minutes.
- Walk-forward splitting with embargo, rolling normalisation, and a configurable cost model are integrated to prevent the most common evaluation pitfalls that inflate reported performance.
- All bootstrap 95% confidence intervals on gross Sharpe ratios include zero; at 15 bps cost, every long-short strategy turns deeply negative—demonstrating the severity of the "alpha cliff."
- After Benjamini–Hochberg FDR correction on 66 pairwise Diebold–Mariano tests, no ML-vs-ML pair is statistically distinguishable.
- Strategy hyperparameters (rebalance frequency, portfolio concentration) shift Sharpe by >0.5—often exceeding model-choice effects—and regime decomposition reveals that COVID-era volatility drives most headline results.