Zhang Yuchen
AI, Ethics and Society Programme
Faculty of Arts
The University of Hong Kong
Hong Kong SAR, China

February 22, 2026

Editor-in-Chief
Engineering Applications of Artificial Intelligence
Elsevier

Dear Editor,

I am writing to submit the manuscript entitled **"A Reproducible Benchmark for Machine Learning in Cross-Sectional Quantitative Trading under Realistic Costs and Regime Shifts"** for consideration as a regular research article in *Engineering Applications of Artificial Intelligence* (EAAI).

Summary. We present **ML Trading Bench**, an open-source evaluation protocol and toolkit for machine learning models applied to cross-sectional quantitative trading. The benchmark addresses a critical gap in the literature: ML trading strategies are routinely evaluated under conditions that inflate reported performance—costs are ignored, splits leak future information, and statistical significance is not tested. Our toolkit integrates walk-forward splitting with embargo, a configurable transaction-cost model, bootstrap confidence intervals, and Diebold–Mariano testing with Benjamini–Hochberg FDR correction into a single pipeline that runs in under 6 minutes with one command.

Key findings. Using 50 US-listed ETFs over 2005–2024, we evaluate 9 models plus 2 passive baselines and demonstrate that: (i) all bootstrap 95% confidence intervals on gross Sharpe ratios include zero; (ii) at 15 bps one-way cost, every long-short strategy turns deeply negative; (iii) after FDR correction, no ML-vs-ML pair is statistically distinguishable; (iv) strategy hyperparameters shift Sharpe by >0.5—more than model choice; (v) regime decomposition reveals that COVID-era volatility drives most headline results.

Relevance to EAAI. This work falls squarely within EAAI's scope of engineering applications of AI. We treat ML-based quantitative trading as an *engineering system evaluation problem*, where protocol design—not model novelty—determines whether conclusions are trustworthy. The paper provides both a practical toolkit (usable by practitioners and researchers) and systematic empirical evidence that evaluation methodology matters as much as model architecture. The complete codebase, data pipeline, and configuration are released under MIT license at `https://github.com/georgekingsman/ml-trading-benchmark`.

Novelty and contribution. Unlike survey papers that *discuss* evaluation pitfalls, our work *demonstrates their consequences* in a concrete, reproducible setting. Unlike trading platforms

(Qlib, FinRL), our focus is on controlled evaluation rather than strategy development. The combination of cost-aware backtesting, regime decomposition, hyperparameter sensitivity analysis, and rigorous statistical testing in a fully automated, one-command pipeline is, to our knowledge, not available in any existing published benchmark.

Declarations. The manuscript has not been published elsewhere and is not under consideration by another journal. All data used are freely available from public sources (no proprietary data). The author declares no conflict of interest. This research did not receive any specific grant funding.

I believe this work will be of significant interest to EAAI's readership in quantitative finance, AI evaluation methodology, and reproducible research. I look forward to your consideration.

Sincerely,

Zhang Yuchen
AI, Ethics and Society Programme
Faculty of Arts, The University of
Hong Kong
u3663696@connect.hku.hk