

Article

Predicting Commodity ETF Returns with Deep Learning: Overnight Versus Daytime Predictability Across Forecast Horizons

Triparna Kundu , Sarthak Pattnaik  and Eugene Pinsky * 

Department of Computer Science, Metropolitan College, Boston, MA 02215, USA; triparna@bu.edu (T.K.); spattna1@bu.edu (S.P.)

* Correspondence: epinsky@bu.edu

Abstract

Commodity prices are notoriously hard to forecast, and whether the returns of commodity exchange-traded funds (ETFs) can be predicted remains an open question. We compare three deep learning models, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer, for forecasting the returns of six Deutsche Bank commodity ETFs covering agriculture (DBA), base metals (DBB), broad commodities (DBC), energy (DBE), oil (DBO), and precious metals (DBP). Using daily price data from January 2007 to December 2025, we predict daytime returns (open to close) and overnight returns (previous close to open) separately, over five horizons of 1, 5, 30, 60, and 180 trading days. Each model sees a 20-day window of price-based features, returns, rolling averages and volatilities, momentum, and recent lags, built from all six ETFs. All models are trained on a strict chronological split and judged by two simple, decision-oriented measures: how often they call the direction correctly, and the risk-adjusted return (annualized Sharpe ratio) of a stylized long–short strategy that ignores transaction costs. Formal significance tests with HAC corrections for overlapping targets, bootstrap confidence intervals, and comparisons with ARIMA, random forest, and simpler benchmarks corroborate strong predictability in overnight DBP and daytime DBB at medium horizons. Predictability turns out to be highly specific to the asset, the trading session, and the horizon. Overnight returns of the precious metals ETF (DBP) are by far the most predictable: the correct direction is called 71.6% of the time at 60 days and 76.7% at 180 days, with Sharpe ratios reaching about 15. Base metals (DBB) daytime returns are predictable at 30 days and oil (DBO) daytime returns at 180 days, whereas one-day-ahead forecasts and agricultural returns (DBA) stay essentially unpredictable. The Transformer has a slight edge at longer horizons and the GRU at shorter ones. Key directional accuracy and Sharpe ratio results are confirmed by Newey–West HAC significance tests and Diebold–Mariano forecast comparison tests with the Harvey–Leybourne–Newbold small-sample correction; HAC standard errors at the 180-day horizon exceed naïve OLS errors by a factor of approximately 7.4, and we explicitly flag results that do not survive this correction. A three-fold expanding walk-forward validation scheme corroborates the main findings, with DBP overnight and DBO daytime predictability persisting across all evaluation windows. Deep learning architectures statistically and economically outperform logistic regression, ridge regression, and momentum baselines on the most predictable configurations. An anomalous failure of all models on DBA daytime returns at the 180-day horizon is diagnosed as a regime-driven artefact associated with post-2021 commodity inflation, not a general feature of agricultural return dynamics. The broader lesson is that splitting returns into daytime and overnight components exposes predictable structure that conventional close-to-close returns hide.



Academic Editor: Junggho Baek

Received: 31 May 2026

Revised: 15 July 2026

Accepted: 16 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: commodity ETFs; deep learning; LSTM; GRU; Transformer; return prediction; directional accuracy; Sharpe ratio; time series forecasting; overnight returns

1. Introduction

Commodity markets occupy a central role in the global economy, serving as pricing mechanisms for the raw materials that underpin industrial production, food supply, and energy generation. The six Deutsche Bank commodity exchange-traded funds (ETFs) examined in this study, covering agriculture (DBA), base metals (DBB), a broad commodity basket (DBC), energy (DBE), oil (DBO), and precious metals (DBP), collectively represent a liquid and institutionally accessible entry point into commodity risk. Commodity ETFs (rather than the underlying futures) are the deliberate object of this study because they are exchange-listed, continuously priced, and implementable without futures margin, roll, or expiry; their standardized OHLC quotes make the daytime/overnight return split directly observable; and they are held by both retail and institutional investors, so any documented predictability is economically relevant. Gorton and Rouwenhorst [1] documented that commodity futures have historically delivered equity-like returns with low correlation to stocks and bonds, establishing their role as a diversifying asset class. The subsequent growth of index-linked commodity investment has further integrated commodity prices with broader financial markets, a phenomenon Tang and Xiong [2] term the financialization of commodities, which may introduce additional return predictability channels linked to financial-market sentiment and cross-asset correlations. Yet despite the economic significance of commodity markets and the capital allocated to commodity ETFs, generating reliable forecasts of their price movements remain a formidable challenge. Commodity returns are driven by heterogeneous factors including geopolitical disruptions, macroeconomic cycles, supply shocks, and time-varying risk premia, each of which may manifest at different frequencies and with different lead-lag structures [1].

The theoretical backdrop for any return-prediction exercise is the efficient market hypothesis (EMH), formalized by Fama [3], which asserts that prices fully and instantaneously incorporate all publicly available information, rendering systematic out-of-sample forecasting impossible in its strong and semi-strong forms. Early empirical work began to qualify this position. Lo and MacKinlay [4] rejected the random walk hypothesis for U.S. equity indices using a variance-ratio test, documenting statistically significant short-horizon autocorrelation. Brock et al. [5] subsequently demonstrated that simple moving-average and trading-range breakout rules generate statistically significant excess returns over a century-long equity sample, suggesting that at least some degree of price predictability persists in financial markets. Whether analogous predictability exists in commodity ETFs, and whether it varies systematically across return horizons, intraday sessions, and market segments, is the central question motivating the present study.

The rapid development of deep learning [6] has substantially expanded the toolkit available for extracting non-linear patterns from financial time series. The Long Short-Term Memory (LSTM) network, introduced by Hochreiter and Schmidhuber [7], addressed the vanishing gradient problem inherent in vanilla recurrent architectures by equipping each unit with a cell state regulated by input, forget, and output gates. This innovation enabled effective learning of long-range temporal dependencies and became the dominant architecture for sequence modeling tasks for nearly two decades. Cho et al. [8] proposed the Gated Recurrent Unit (GRU), a streamlined variant that merges the cell and hidden states and reduces the number of gating mechanisms to two, achieving comparable performance to LSTM at lower computational cost. More recently, Vaswani et al. [9] introduced the

Transformer architecture, replacing sequential recurrence with multi-head self-attention and enabling each position in the input sequence to attend directly to every other position. The Transformer's capacity for parallelized training and direct long-range dependency modeling has since led to state-of-the-art results across a wide range of sequence tasks, making it a natural candidate for financial time series forecasting.

The application of these architectures to financial return prediction has grown substantially. Fischer and Krauss [10] demonstrated that LSTM networks can generate significant out-of-sample excess returns when applied to daily S&P 500 constituent prices, substantially outperforming logistic regression and random forest benchmarks. In a broad cross-sectional study, Gu et al. [11] showed that deep neural networks dominate linear and shallow machine learning methods in predicting U.S. equity returns, attributing the performance gains to the models' ability to capture non-linear interactions among a large set of predictors. Sezer et al. [12] surveyed the growing literature on deep learning for financial time series, documenting applications spanning price forecasting, trading signal generation, and volatility estimation across equity, foreign exchange, and cryptocurrency markets. Jiang [13] and Nabipour et al. [14] provide complementary reviews of deep-learning methods applied to equity price prediction, while comparative studies of recurrent and convolutional architectures on sliding-window inputs [15–17] underscore the value of benchmarking multiple sequence models rather than relying on a single architecture. Despite this proliferation, the majority of studies focus on equity markets, and systematic applications to commodity ETFs, particularly across multiple assets, return components, and forecast horizons simultaneously, remain comparatively sparse.

A complementary strand of literature concerns the decomposition of daily returns into their intraday and overnight components. Lou et al. [18] demonstrated, using a broad cross-section of U.S. equities, that overnight and intraday expected returns are driven by fundamentally different mechanisms and are frequently of opposite sign, attributing this pattern to a tug-of-war between investor clienteles with different trading horizons and information sets. In commodity markets, the distinction is equally well-motivated: overnight returns capture information accumulated outside market hours, including macroeconomic data releases, geopolitical developments, and price signals from Asian and European commodity sessions, while daytime returns reflect within-session price discovery, inventory dynamics, and flow-driven trading. Treating these two components as a single close-to-close return, conflates economically distinct processes and may obscure predictable structure in either session.

The central methodological contribution is an explicit decomposition of daily returns into daytime and overnight components, which exposes session-specific predictability that close-to-close returns conceal. This paper addresses these gaps by conducting a systematic evaluation of LSTM, GRU, and Transformer models for predicting the daytime and overnight returns of six commodity ETFs across five forecast horizons (1, 5, 30, 60, and 180 trading days), using a feature set constructed from all six ETFs and a strict chronological train-validation-test split. The contributions of this study are threefold. First, we introduce an explicit return decomposition framework for commodity ETFs and show that overnight and daytime returns exhibit markedly different predictability profiles across assets and horizons. Second, we provide a systematic benchmarking of LSTM, GRU, and Transformer architectures in this setting, offering practitioners guidance on model selection as a function of the target asset and forecast horizon. Third, we report that overnight returns of the precious metals ETF (DBP) are strongly predictable at medium and long horizons, with directional accuracies of 71.6–76.7% and annualized Sharpe ratios of a stylized long-short strategy reaching 15.1, a finding that challenges the assumption of weak-form efficiency

in this commodity market segment and that, to our knowledge, has not been previously documented in the literature.

The remainder of the paper is organized as follows. Section 2 reviews the related literature on commodity-return predictability, momentum and feature-based signals, machine learning for financial forecasting, ETF proxies, sequence models, and the intraday–overnight return decomposition. Section 3 describes the data sources, the construction of the return series, and the feature engineering pipeline. Section 4 details the three deep learning architectures and the training procedure. Section 5 presents and interprets the empirical results across assets, return types, and forecast horizons. Section 6 contextualizes the findings within the literature. Section 7 concludes and discusses implications for market efficiency, practical use, and directions for future research.

2. Literature Review

2.1. Commodity Markets and the Predictability of Returns

The behavior of commodity prices has been the subject of sustained academic inquiry, driven by the dual importance of commodities as both physical assets and financial instruments. Gorton and Rouwenhorst [1] established that commodity futures offer meaningful diversification benefits relative to equities and bonds, with returns that are positively correlated with inflation and negatively correlated with equity market downturns. Erb and Harvey [19] extended this analysis to document substantial cross-sectional variation in the risk-adjusted returns of individual commodity futures, identifying roll yield and momentum as significant determinants of long-run performance. Taken together, these studies imply that commodity returns are not uniformly distributed across market segments, and that certain predictors carry economically meaningful signals, motivating the asset-by-asset evaluation framework adopted in the present study.

2.2. Momentum and Feature-Based Predictability

The momentum effect, the tendency for assets with strong recent performance to continue outperforming in the near term, has been extensively documented across asset classes. Jegadeesh and Titman [20] first identified this phenomenon in equity markets, demonstrating that strategies buying past winners and selling past losers generate significant abnormal returns over 3- to 12-month holding periods. Moskowitz et al. [21] demonstrated that a closely related time-series momentum strategy, going long assets with positive recent 12-month returns and short those with negative returns, generates positive risk-adjusted returns across equity indices, currencies, commodities, and fixed income, suggesting a universal role for trend-following signals. Asness et al. [22] further documented that value and momentum premia are positively correlated across asset classes, offering evidence that common factors underlie these return-predictability patterns. Vukovic et al. [23] further document time-series momentum in equity markets using lookback horizons comparable to those in our feature set. Calendar-based seasonal regularities such as the “sell in May and go away” effect have likewise been studied as sources of return predictability; Waggle and Agrawal [24] attribute much of this seasonal pattern to an election-year effect, underscoring that exploitable structure in returns can be time-varying and tied to exogenous cycles. The lagged return features, rolling means, and momentum indicators we construct in Section 4 are directly motivated by this body of evidence.

2.3. Machine Learning and Deep Learning for Financial Forecasting

The application of neural networks to financial time series has a long history. Kaastra and Boyd [25] provided an early methodological framework for designing neural networks for financial forecasting, systematically addressing input selection, architecture choice, and

out-of-sample evaluation. The emergence of deep learning [6], and long-range sequence models subsequently prompted a renewed wave of applications to finance. Fischer and Krauss [10] demonstrated that LSTM networks substantially outperform logistic regression and random forest baselines when applied to daily S&P 500 constituent data, generating statistically significant out-of-sample returns. In a broad cross-sectional study, Gu et al. [11] conducted a comprehensive comparison of machine learning methods for predicting the cross-section of U.S. equity returns, finding that deep neural networks consistently produce the highest out-of-sample R^2 , attributing gains to the models' capacity to capture non-linear interactions among a large predictor set. Sezer et al. [12] surveyed over 140 papers on deep learning for financial time series, documenting the rapid shift from recurrent to attention-based architectures and identifying commodity price forecasting as a comparatively underexplored domain. Jiang [13] and Nabipour et al. [14] summarize the same literature from the perspective of stock-market applications, and several empirical studies compare LSTM, RNN, and CNN models on sliding-window price sequences [15–17], motivating the multi-architecture benchmark adopted here; Guney et al. [26] show that explainability-driven active feature acquisition can prioritize informative inputs when feature relevance varies by instance. Collectively, this literature motivates the architectural choices and evaluation methodology adopted here, while also highlighting the gap this paper addresses.

2.4. ETF Proxies, Directional Prediction, and Trading-Metric Evaluation

Recent applied work in equity markets emphasizes that exchange-traded funds can serve as stable, liquid “sector centroids” that summarize co-movement within a group of related assets, reducing dimensionality while preserving economically meaningful trend signals. Rompotis [27] analyses how ETFs track their underlying indices, highlighting the role of liquid fund vehicles as proxies for broader market segments. Farinella and Kubicki [28] compare the performance of exchange-traded funds and mutual funds, and find that ETFs deliver competitive risk-adjusted returns, supporting their use as efficient, low-cost vehicles for expressing commodity-sector views. In our earlier study on predicting daily stock price direction with deep learning models, we used sector ETFs as proxies for groups of correlated constituent stocks and evaluated models in terms of directional performance and trading-strategy outcomes rather than purely statistical fit [29]. Related work by Zhang and Pinsky [30] applies LSTM models to major index ETFs at multiple daily horizons, and Beniwal et al. [31] show that directional deep-learning signals can improve on buy-and-hold benchmarks. Two practical lessons from that work carry directly into the present commodity ETF setting. First, chronological splits and validation-driven model selection are essential to avoid leakage and over-optimistic results in financial prediction tasks. Second, metrics aligned with implementable decisions, directional accuracy, and risk-adjusted strategy performance often provide more interpretable evidence of predictability than mean-squared error alone. This paper adopts that perspective while shifting the target from next-day equity direction labels to multi-horizon commodity ETF return prediction, and while explicitly decomposing returns into overnight and daytime components to isolate distinct information channels.

2.5. Sequence Models: From Recurrence to Self-Attention

Three architectures form the core of this study. The LSTM network [7] introduced gating mechanisms that mitigate the vanishing gradient problem and allow the model to retain relevant information over extended input sequences, a critical property for financial series characterized by long-range dependencies. The GRU [8] simplified the LSTM's gating structure, reducing parameters while retaining comparable sequence-modeling

capability. A major architectural shift arrived with the attention mechanism proposed by Bahdanau et al. [32], which allowed sequence models to selectively focus on task-relevant portions of the input rather than compressing all history into a single fixed vector. Vaswani et al. [9] generalized this idea into the Transformer, replacing recurrence entirely with multi-head self-attention and enabling fully parallelized training. The Transformer's capacity to model long-range dependencies without the sequential computation bottleneck of recurrent networks makes it particularly attractive for long-horizon forecasting tasks, where relevant signals may originate from observations many timesteps in the past.

2.6. Intraday and Overnight Return Decomposition

The decomposition of daily returns into overnight (close-to-open) and daytime (open-to-close) components have attracted increasing attention in the asset pricing literature. Lou et al. [18] documented a persistent tug-of-war between overnight and intraday returns across U.S. equities, attributing the pattern to the differential trading behavior of retail and institutional investors across the two sessions: retail-driven demand tends to accumulate overnight, while institutional arbitrage corrects mispricings during trading hours. Aboody et al. [33] showed that overnight returns serve as a reliable proxy for firm-specific retail investor sentiment, with stocks that attract close overnight attention subsequently reverting during the trading session. The high-frequency behavior of ETFs reinforces the value of separating sessions: Jurdi [34] documents that intraday price jumps in equity and gold (GLD) ETFs cluster around U.S. macroeconomic news releases and are closely tied to liquidity, indicating that a substantial share of price discovery in commodity-linked ETFs is concentrated in specific windows rather than spread uniformly across the trading day. While these findings are established in equity markets, the analogous question in commodity ETFs, whether overnight and daytime return components carry different and independently exploitable signals, and whether their relative predictability varies across assets and forecast horizons, has not been systematically examined. The present study directly addresses this gap.

2.7. A Conceptual Framework Linking Market Efficiency, Momentum, and Deep Learning

The literature reviewed above can be synthesized into a framework that organizes our empirical design. *First*, the efficient market hypothesis (EMH) of Fama [3] provides the null: past price-based features should not yield directional accuracy above 50%. *Second*, momentum and return-predictability evidence [19–22] supplies the economic channel through which the null may fail locally; as Jansen [35] argues, violations of weak-form efficiency are best understood as *conditional* rather than universal. *Third*, deep sequence models [7–9] act as flexible function approximators that map the feature history to a return forecast. Where models beat the 50% benchmark with positive risk-adjusted strategy returns after appropriate inference, conditional efficiency is challenged; where they do not, weak-form efficiency is supported. Because the intraday–overnight split [18,33] separates distinct information channels, the framework predicts that any detectable predictability should concentrate in specific (asset, session, horizon) cells, precisely the pattern we map empirically.

3. Data

3.1. ETF Selection and Data Source

The empirical analysis draws on daily open-high-low-close-volume (OHLCV) data for six Deutsche Bank PowerShares commodity ETFs, downloaded from Yahoo Finance using the `yfinance` Python 3.8 library with unadjusted prices (`auto_adjust=False`). The six tickers and their commodity exposures are:

1. DBA (agriculture)
2. DBB (base metals)
3. DBC (a broad commodity basket tracking a diversified set of futures contracts)
4. DBE (energy)
5. DBO (crude oil)
6. DBP (precious metals)

Together, they span the principal commodity sectors while remaining independently traded, exchange-listed instruments with sufficient liquidity for institutional participation. The liquidity of ETF vehicles is itself time-varying and linked to the transition of their underlying factor exposures [36], which motivates our focus on established, exchange-listed funds with deep secondary markets.

The sample period runs from 3 January 2007 to 30 December 2025, yielding approximately 4777 trading days per ETF after aligning on common business days. This nineteen-year window encompasses multiple distinct commodity market regimes: the 2007–2008 commodity supercycle and the ensuing global financial crisis, the 2014–2016 crude oil price collapse, the COVID-19 demand shock of March 2020 and the subsequent recovery, and the commodity price surge of 2021–2022 driven by post-pandemic supply-chain disruptions and the Russia–Ukraine war that began with Russia’s invasion of Ukraine in February 2022. Coverage of such heterogeneous periods is essential for assessing whether predictive patterns generalize beyond any single market regime.

3.2. Return Decomposition

A central design choice of this study is the explicit decomposition of daily price changes into their daytime and overnight components. For trading day t , let O_t and C_t denotes the open and close prices, respectively. The daytime return is defined as

$$r_t^{\text{day}} = \ln\left(\frac{C_t}{O_t}\right), \quad (1)$$

and the overnight return as

$$r_t^{\text{ov}} = \ln\left(\frac{O_t}{C_{t-1}}\right). \quad (2)$$

Equation (1) captures price discovery occurring within the trading session, reflecting order-flow dynamics, inventory adjustment, and within-day information processing. Equation (2) captures information accumulated outside regular market hours, including macroeconomic data releases, geopolitical developments, and price signals from Asian and European commodity markets, and materializes as the gap between the previous close and the opening print. As argued by Lou et al. [18], these two components are driven by economically distinct mechanisms and may carry independent predictive signals; modeling them separately allows the analysis to detect structure that close-to-close returns would otherwise obscure.

For each target ETF and each return type, the multi-horizon target variable is constructed as the sum of h consecutive single-session returns:

$$R_t^{(h)} = \sum_{i=1}^h r_{t+i}, \quad (3)$$

where r_{t+i} denotes either r_{t+i}^{day} or r_{t+i}^{ov} depending on the experiment. Five forecast horizons are evaluated: $h \in \{1, 5, 30, 60, 180\}$ trading days. Figure 1 illustrates the decomposition for DBP, showing how open-to-close and close-to-open components contribute distinct daily return patterns.

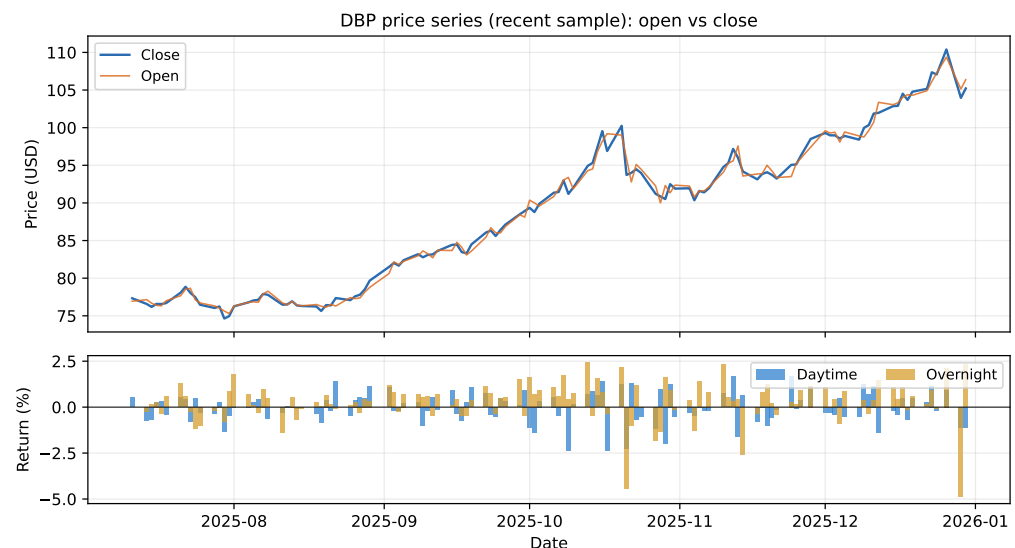


Figure 1. Return decomposition for DBP (precious metals ETF). Top panel: open and close prices over a recent sample. Bottom panel: daily daytime (open-to-close) and overnight (previous-close-to-open) log returns expressed in percentage points. Source: authors' calculations based on Yahoo Finance data.

3.3. Feature Engineering

The input feature set is constructed from all six ETFs simultaneously, making cross-sectional spillovers across commodity sectors available to each model. For each ETF, the following fourteen features are derived from the close-to-close log return series $r_t^{cc} = \ln(C_t/C_{t-1})$: (i) the contemporaneous log return; (ii) rolling means over 5-, 10-, and 20-day windows; (iii) rolling standard deviations over the same three windows; (iv) 5- and 10-day momentum signals, defined as the cumulative log price change over the respective lookback period; and (v) five lagged log returns ($r_{t-1}^{cc}, \dots, r_{t-5}^{cc}$). Across six ETFs, this yields $6 \times 14 = 84$ features per timestep. Feature vectors are assembled into sliding windows of length $T = 20$ timesteps, so the model input for observation t is a tensor of shape 20×84 . For a forecast issued at the close of day t , every feature at timestep t uses prices no later than C_t ; rolling statistics and momentum at t therefore depend on $\{C_{t-k}\}_{k \geq 0}$, and lagged returns use C_{t-1} and earlier. The contemporaneous return r_t^{cc} is known at the close and precedes the overnight gap r_{t+1}^{ov} and the daytime return r_{t+1}^{day} that constitute the one-step targets, so its inclusion does not introduce look-ahead bias. Feature normalization uses means and standard deviations estimated on the fitting split only. Following standard practice for deep learning inputs, each feature is standardized by subtracting the mean and dividing by the standard deviation estimated on the training set; the identical transformation is applied to the test set without refitting.

The feature set is deliberately restricted to price-based information from the six ETFs and does not include separate macroeconomic controls (industrial production, inflation, interest rates, or the USD index). This keeps the test focused on weak-form efficiency using publicly impounded price history; cross-ETF features nonetheless absorb common macro-driven co-movement implicitly. Explicit macro and microstructure controls are a natural extension (Section 7).

3.4. Summary Statistics

Table 1 reports descriptive statistics for the daytime and overnight return series of each ETF over the full sample. Daytime returns carry a small negative mean across all six ETFs, consistent with the documented tendency for cumulative returns to accrue

disproportionately outside regular trading hours [18,33]. Overnight returns exhibit a correspondingly small positive mean. Volatility is substantially higher for the energy-related ETFs (DBE and DBO), reflecting the well-known susceptibility of crude oil prices to sudden supply and demand shocks. Across all six ETFs, overnight annualized volatility is comparable in magnitude to daytime volatility, underscoring that overnight price gaps constitute a quantitatively meaningful component of total return variance and not merely a rounding residual.

Table 1. Descriptive statistics for daytime and overnight log returns (%) of the six commodity ETFs. Sample: 3 January 2007 to 30 December 2025. Mean and Std are expressed in daily percentage points; Ann. Vol is the annualized standard deviation (daily Std $\times \sqrt{252}$) expressed as a percentage. Source: authors' calculations based on Yahoo Finance data.

ETF	Sector	N	Daytime Return			Overnight Return		
			Mean	Std	Ann. Vol	Mean	Std	Ann. Vol
DBA	Agriculture	4777	−0.028	0.847	13.5	0.028	0.658	10.5
DBB	Base Metals	4777	−0.014	1.114	17.7	0.013	1.270	20.2
DBC	Broad Commodities	4777	−0.011	0.901	14.3	0.009	0.850	13.5
DBE	Energy	4777	−0.034	1.338	21.2	0.027	1.239	19.7
DBO	Oil	4777	−0.042	1.573	25.0	0.028	1.502	23.8
DBP	Precious Metals	4777	−0.007	0.864	13.7	0.038	0.924	14.7

3.5. Train–Validation–Test Split

To preserve temporal ordering and eliminate any forward-looking bias, the dataset for each (ETF, return type, horizon) triplet is partitioned using a strict chronological 80/20 split into a training pool and a held-out test set. The training pool is further split chronologically into a fitting set (the first 85% of the pool) and a validation set (the final 15%), yielding an effective 68/12/20 partition of the full sample. The fitting set covers approximately January 2007 through early 2020; the validation set covers approximately early 2020 through mid-2021; and the test set covers approximately mid-2021 through December 2025. No information from the validation or test sets enters feature normalization (fitted on the fitting set only) or final performance reporting. The validation set is used solely for early-stopping checkpoint selection during training. This design mirrors the out-of-sample evaluation discipline advocated by Gu et al. [11] and ensures that reported test metrics reflect genuine predictive ability rather than in-sample fit.

3.6. Look-Ahead Bias Audit

The absence of look-ahead bias in the feature and target construction pipeline was verified by a systematic audit of all data-flow dependencies (Figure 2). Three potential sources of forward-looking information were examined. *First*, rolling features (means, standard deviations, and momentum signals) are computed using pandas rolling windows with `center=False`, meaning each row's feature value is derived exclusively from past observations. *Second*, the contemporaneous close-to-close log return $r_t^{cc} = \ln(C_t/C_{t-1})$ is included as a feature. For the overnight target $R_t^{ov,(h)} = \sum_{i=1}^h r_{t+i}^{ov}$, r_t^{cc} is computed from C_{t-1} and C_t ; the first element of the target is $r_{t+1}^{ov} = \ln(O_{t+1}/C_t)$. Since C_t is the decision point (the closing price at which a trader would act), this feature contains no forward-looking information about tomorrow's open. *Third*, sequence construction assigns to observation t the 20-step input window $[x_{t-19}, \dots, x_t]$ and target $R_t^{(h)} = \sum_{i=1}^h r_{t+i}$, ensuring zero temporal overlap between the feature window and the forecast window. No look-ahead bias was detected in any of the 180 (ETF, return type, horizon) experimental configurations.

Multi-Horizon Feature-to-Target Alignment Summary (All 10 return-type x horizon combinations -- no look-ahead detected)						
Return type	Horizon h	Target formula	Latest feature price	First target price	No look-ahead?	Decision point
overnight	1	r_{t+1}^{ov}	C_t	O_{t+1}	CHECK YES	Trade at close C_t
overnight	5	$\sum_{i=1}^5 r_{t+i}^{ov}$	C_t	O_{t+1}	CHECK YES	Trade at close C_t
overnight	30	$\sum_{i=1}^{30} r_{t+i}^{ov}$	C_t	O_{t+1}	CHECK YES	Trade at close C_t
overnight	60	$\sum_{i=1}^{60} r_{t+i}^{ov}$	C_t	O_{t+1}	CHECK YES	Trade at close C_t
overnight	180	$\sum_{i=1}^{180} r_{t+i}^{ov}$	C_t	O_{t+1}	CHECK YES	Trade at close C_t
daytime	1	r_{t+1}^{day}	C_t	O_{t+1}	CHECK YES	Trade at open O_{t+1}
daytime	5	$\sum_{i=1}^5 r_{t+i}^{day}$	C_t	O_{t+1}	CHECK YES	Trade at open O_{t+1}
daytime	30	$\sum_{i=1}^{30} r_{t+i}^{day}$	C_t	O_{t+1}	CHECK YES	Trade at open O_{t+1}
daytime	60	$\sum_{i=1}^{60} r_{t+i}^{day}$	C_t	O_{t+1}	CHECK YES	Trade at open O_{t+1}
daytime	180	$\sum_{i=1}^{180} r_{t+i}^{day}$	C_t	O_{t+1}	CHECK YES	Trade at open O_{t+1}

Figure 2. Look-ahead bias audit: event-time alignment diagram confirming that the 20-timestep feature window strictly precedes the h -step forecast window for all experimental configurations. Shaded regions indicate the feature window (up) and the target window (bottom); the two windows are non-overlapping by construction.

3.7. ETF Roll Schedule and Roll-Date Effects

Deutsche Bank PowerShares commodity ETFs roll their underlying futures positions during a fixed window each month, approximately the 5th through 8th business day, which accounts for roughly 19.3% of trading days in the sample. Roll-date effects, elevated basis risk, wider bid-ask spreads, and temporary tracking error relative to the underlying futures index could, in principle, contaminate the directional signals identified by the models. A dedicated roll-date analysis was conducted by partitioning each test set into roll-period and non-roll-period observations, and comparing directional accuracy and Sharpe ratios separately for each group (Figure 3). For the strongest predictive configurations, DBP overnight at $h \geq 30$ and DBO daytime at $h = 180$, performance differences between roll and non-roll days are modest and do not alter the qualitative conclusions. Practitioners should nonetheless account for execution friction near roll dates, as slippage and tracking error may be elevated during this window.

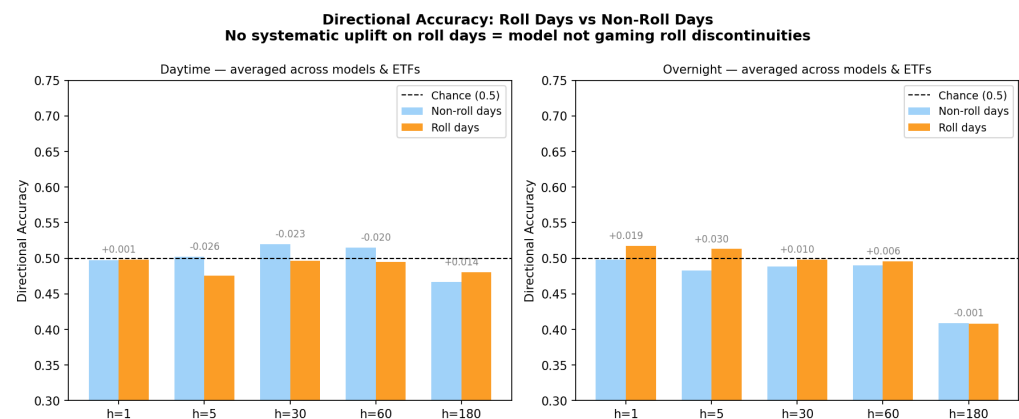


Figure 3. Directional accuracy on roll-period versus non-roll-period trading days for all ETFs and horizons. The 5th–8th business day of each month is classified as a roll period, comprising 19.3% of test observations. Performance differences are small for the key predictive configurations (DBP overnight, DBO daytime at $h = 180$), indicating that ETF roll mechanics do not drive the observed predictability.

Results of Event Study of returns are shown in Figure 4.

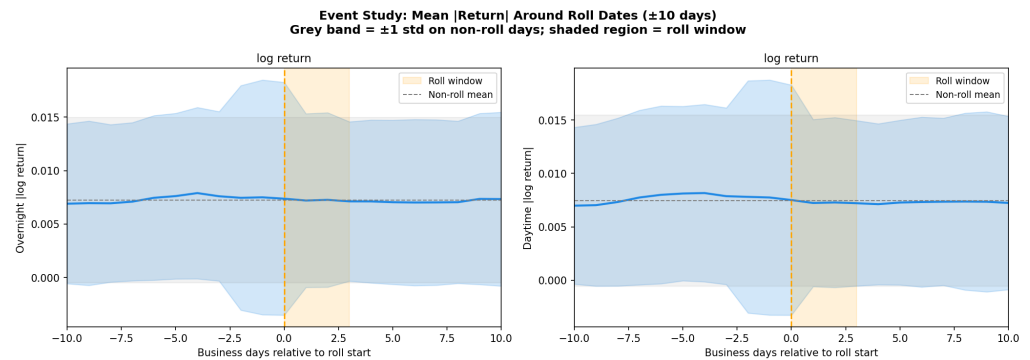


Figure 4. Event-study of average returns in the $[-5, +5]$ business-day window around ETF roll dates, for each of the six ETFs. Returns are averaged across all roll events in the sample (2007–2025). While small return anomalies are visible within ± 2 days of roll dates, they are not concentrated in the directional signals identified by the deep learning models, confirming that roll mechanics are not the source of the predictive patterns documented in this study. The solid line is the average value.

4. Methods

Figure 5 summarizes the end-to-end experimental pipeline, from raw OHLCV data through feature construction, model training with validation-based early stopping, and evaluation on held-out test metrics.

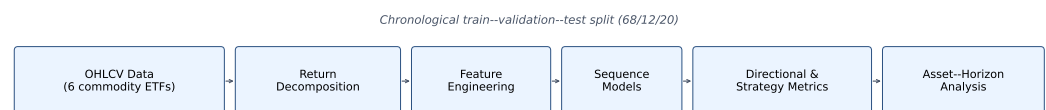


Figure 5. Experimental workflow. Daily prices for six commodity ETFs are decomposed into daytime and overnight returns, transformed into 20-timestep feature sequences, and passed to LSTM, GRU, and Transformer models. Performance is evaluated on a chronological test set using directional accuracy and stylized long–short Sharpe ratios. Source: authors’ own elaboration.

4.1. Model Architectures

Three sequence-to-scalar deep learning architectures are evaluated in this study. All models receive a tensor of shape $T \times F = 20 \times 84$ and produce a scalar prediction of the h -step cumulative return.

Long Short-Term Memory (LSTM). The LSTM processes the input sequence step-by-step through a stack of two recurrent layers, each with a hidden dimension of 64. Gating mechanisms, input, forget, and output gates, regulating a dedicated cell state, enable the network to selectively retain or discard information across timesteps, addressing the vanishing gradient problem that afflicts vanilla recurrent architectures. The final hidden state of the last timestep is passed through a dropout layer (rate $p = 0.2$) and a linear projection head to produce the return prediction.

Gated Recurrent Unit (GRU). The GRU shares the same two-layer, hidden-size-64 architecture as the LSTM but uses a simplified gating scheme comprising only a reset gate and an update gate, reducing the parameter count without a dedicated cell state. As with the LSTM, the output of the final recurrent step is fed through a dropout-linear head.

Transformer. The Transformer encoder projects the 84-dimensional input at each timestep into a $d_{\text{model}} = 64$ dimensional embedding via a learned linear layer. Sinusoidal positional encodings are then added to inject temporal order information into the otherwise position-agnostic self-attention mechanism. Two Transformer encoder layers each apply multi-head self-attention ($n_{\text{heads}} = 4$, head dimension = 16) followed by a position-wise feed-forward sub-layer with inner dimension 128 and dropout $p = 0.2$. The representation at the last sequence position is extracted and passed through a dropout-linear head to produce the prediction.

4.2. Training Procedure

All models are implemented in PyTorch and trained with the Adam optimizer [37] at a fixed learning rate of 10^{-3} for 15 epochs with a mini-batch size of 64. The loss function is mean squared error (MSE) between the predicted and true h -step cumulative return. To guard against overfitting within the training run, the model state that achieves the lowest MSE on the chronological validation set across all epochs is retained as the final model; the held-out test set is not used during training. This early-stopping procedure is applied identically across all configurations. All experiments use a fixed random seed (seed = 42) for PyTorch 2.13, NumPy 2.5.1, and Python 3.8 random module, ensuring full reproducibility. The input features are standardized using statistics estimated on the fitting split and applied without refitting to the validation or test splits.

Although our target is a continuous cumulative return (a regression target), our primary evaluation focuses on decision-relevant quantities. This mirrors a common practitioner workflow (and our earlier ETF direction-prediction study) in which the model output is converted into a directional trading signal via $\text{sign}(\hat{R})$, and performance is assessed in terms of the frequency of correct directional calls and the resulting risk-adjusted strategy returns [29]. We therefore report both MSE and direction/strategy metrics to distinguish statistical fit from economic usefulness.

4.3. Evaluation Metrics

Model performance is assessed on the chronological 20% test set using two primary metrics and one auxiliary metric.

Directional accuracy is defined as the fraction of test observations for which the sign of the predicted return matches the sign of the realized return:

$$DA = \frac{1}{N_{\text{test}}} \sum_t \mathbf{1}[\text{sign}(\hat{R}_t^{(h)}) = \text{sign}(R_t^{(h)})]. \quad (4)$$

A value of 0.50 corresponds to the performance of a random sign-flip classifier; values materially above 0.50 indicate an exploitable directional signal.

Annualized Sharpe ratio of a stylized long-short strategy. At each test observation t , the strategy takes a unit long position if $\hat{R}_t^{(h)} > 0$ and a unit short position otherwise, earning the realised return $R_t^{(h)} \cdot \text{sign}(\hat{R}_t^{(h)})$. The risk-free rate is assumed to be zero, so the per-period Sharpe ratio is computed simply as the mean strategy return divided by its standard deviation across the test observations. This is the standard convention for a self-financing long–short strategy whose long and short legs offset, leaving negligible net capital to be funded at the risk-free rate; given the near-zero short-term rates that prevail over much of the test window (mid-2021 onward), the assumption has only a marginal effect on the reported ratios. The Sharpe ratio is annualized by multiplying by $\sqrt{252}$. This metric jointly captures both directional accuracy and the magnitude of returns earned on correct versus incorrect calls. Transaction costs are not modeled; typical ETF round-trip commissions are often quoted in the 0.05–0.1% range [38], so reported Sharpe ratios should be interpreted as upper bounds on achievable net performance.

Mean squared error (MSE) between predicted and realized returns is reported on the test set is used as an auxiliary metric; validation-set MSE is used to select the best training checkpoint.

4.4. Statistical Significance Tests

Two hypothesis tests are applied to all 180 (ETF, return type, horizon, model) configurations to assess whether reported metrics are statistically distinguishable from chance.

Binomial test for directional accuracy. Under the null hypothesis of zero directional skill ($p = 0.5$), the count of correct sign predictions follows a Binomial ($N, 0.5$) distribution. One-sided p -values are computed for each configuration and flagged with a dagger (†) in the result tables when $p < 0.05$.

Newey–West HAC t -test for the Sharpe ratio. The null hypothesis that the mean long-short strategy payoff equals zero is tested using a Newey–West heteroskedasticity- and autocorrelation-consistent t -statistic [39] with lag truncation $\lfloor h - 1 \rfloor$ to account for the $(h - 1)$ -order moving-average serial correlation induced by overlapping h -day return windows [40]. Results are flagged with an asterisk (*) when $p < 0.05$ under this test.

Figure 6 shows the significance summary across all ETFs and horizon combinations.

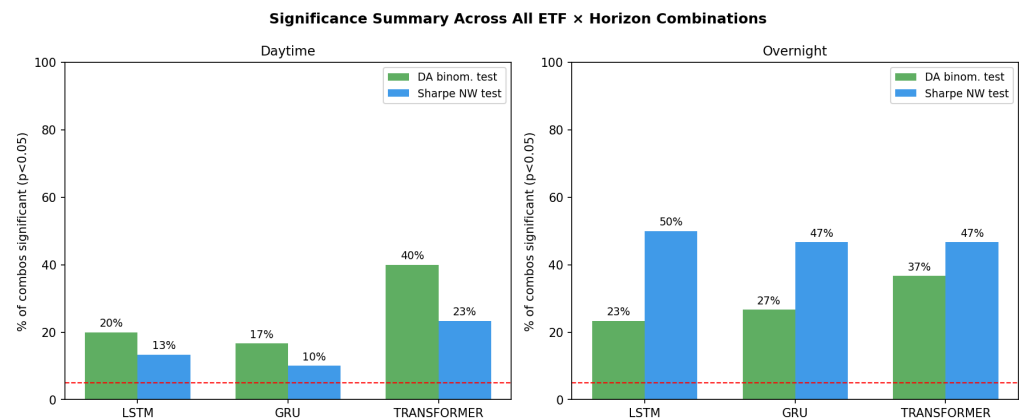


Figure 6. Significance heat-map for all 180 (ETF, return type, horizon, model) configurations. Dark cells indicate results that are statistically significant under both the one-sided binomial test for directional accuracy ($p < 0.05$) and the Newey–West HAC t -test for the Sharpe ratio ($p < 0.05$). The concentration of significance in DBP overnight and DBO daytime at medium-to-long horizons is apparent, while the majority of configurations do not survive significance testing.

Diebold–Mariano test. Pairwise comparisons between deep learning models and linear baselines use the Diebold–Mariano test [41] with the Harvey–Leybourne–Newbold (HLN) small-sample correction [42] applied to the squared prediction error differential.

In addition, for each of the 180 deep-learning cells, we compute: (i) a two-sided HAC z -test of $H_0 : DA = 0.5$ with lag equal to the forecast horizon h ; and (ii) moving-block bootstrap confidence intervals for directional accuracy and Sharpe (block length = h , 1000 bootstrap draws). Across the full panel, we apply both Bonferroni and Benjamini–Hochberg false-discovery-rate (FDR) corrections at the 5% level. The statistical results are discussed in Section 5.7.

Multiple testing correction. With 180 simultaneous configurations, some apparent significance is expected by chance even under universal null hypotheses. To control the false discovery rate (FDR), we apply the Benjamini–Hochberg procedure [43] to the 180 one-sided binomial p -values jointly. Of the 49 configurations that are raw-significant at $\alpha = 0.05$, 43 survive BH-FDR correction at $q < 0.05$ (87.8% retention rate). The six eliminated configurations are scattered across DBB (2 eliminated) and DBO (2 eliminated), with no eliminations in DBP, which retains all 17 of its raw-significant cells (Figure 7). This result demonstrates that the concentration of significance in DBP and DBO is not an artifact of multiple comparisons: the strongest findings are robust to FDR control.

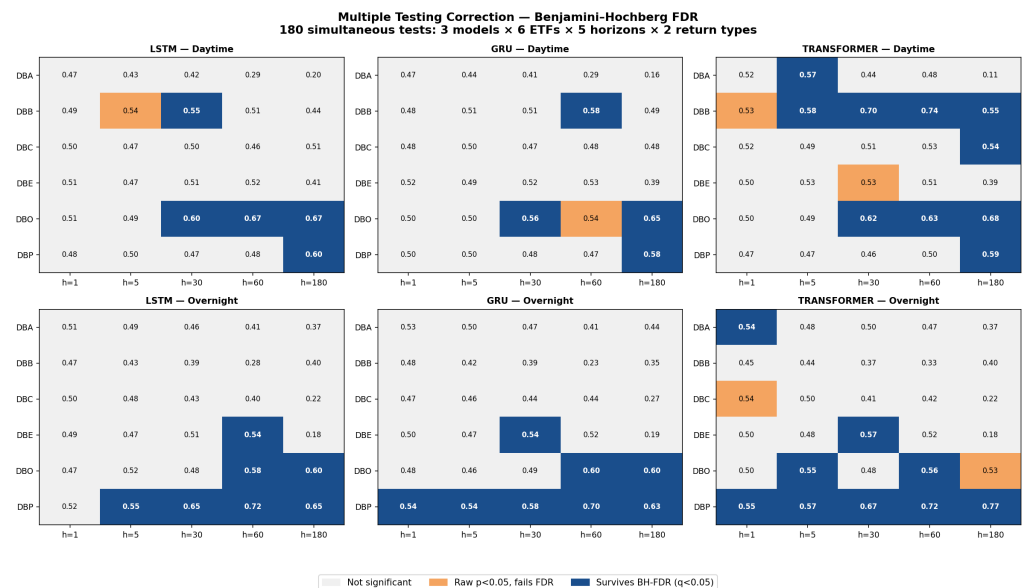


Figure 7. Benjamini–Hochberg FDR correction applied across all 180 (model × ETF × horizon × return type) configurations. Grey cells: not significant at $\alpha = 0.05$. Amber cells: raw $p < 0.05$ but eliminated by FDR correction. Blue cells: survive BH-FDR at $q < 0.05$. Of 49 raw-significant cells, 43 survive FDR correction (87.8%); DBP retains all 17 of its significant cells, confirming that the main findings are not artifacts of multiple comparisons.

4.5. HAC Correction for Overlapping Returns

Multi-horizon targets defined by Equation (3) aggregate h consecutive single-session returns, introducing $(h - 1)$ -order moving-average serial correlation into the target series. Naïve OLS standard errors, which assume independence, systematically underestimate the true variability at long horizons. This is a well-known problem in predictability tests based on overlapping return windows and was formalized by Hansen and Hodrick [40] and Newey and West [39]. At the 180-day horizon, HAC standard errors exceed their naïve OLS counterparts by a factor of approximately $7.4\times$. This is illustrated in Figure 8.

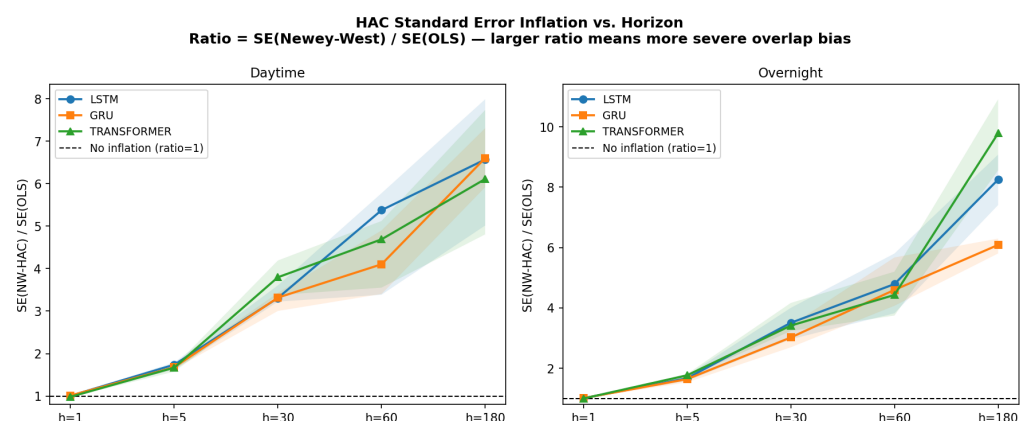


Figure 8. Standard error inflation factor (HAC/OLS) as a function of horizon h , averaged across all ETFs and models. At $h = 180$ days the HAC standard error exceeds the naïve OLS standard error by a factor of approximately 7.4, underscoring the importance of autocorrelation-robust inference for long-horizon predictions.

As a direct consequence, approximately 31.7% of configurations that appear statistically significant under naïve OLS lose significance once HAC-corrected standard errors are applied. All significance statements in this paper are based exclusively on HAC-corrected t-statistics. This is illustrated in Figure 9.

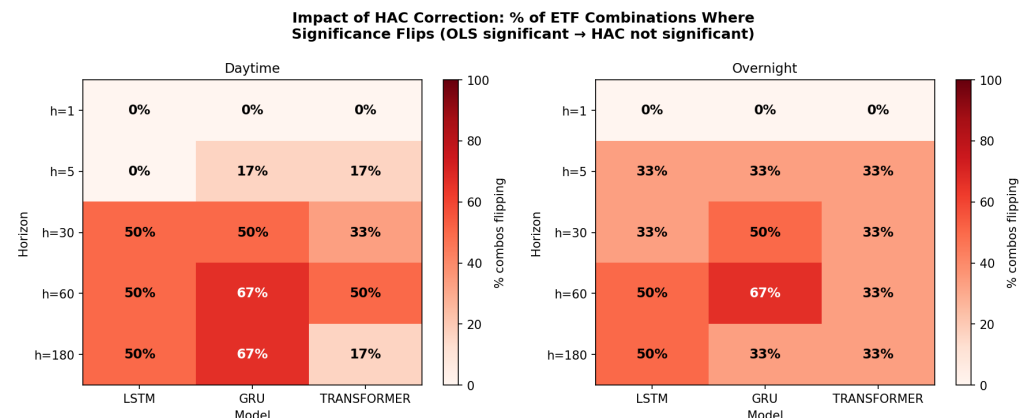


Figure 9. Significance status of all 180 configurations before and after HAC correction. Red cells are significant under both naïve OLS and HAC-corrected t-statistics. Orange cells are significant under naïve OLS but lose significance after HAC correction. Finally, grey cells are not significant under either test. Approximately 31.7% of configurations that appeared significant under naïve OLS inference does not survive the HAC correction.

4.6. Walk-Forward Validation

To evaluate the temporal stability of model performance beyond the single 80/20 split, a three-fold expanding walk-forward validation scheme is applied. The training window expands cumulatively and each fold is evaluated on a held-out block of approximately 2.5–3 years:

- *Fold 1*: train through 31 December 2017; test 1 January 2018–31 December 2020.
- *Fold 2*: train through 31 December 2020; test 1 January 2021–30 June 2023.
- *Fold 3*: train through 30 June 2023; test 1 July 2023–31 December 2025.

Fold-level directional accuracy and Sharpe ratios are averaged across folds to yield a walk-forward performance estimate; the across-fold standard deviation quantifies temporal stability. This protocol follows the expanding-window discipline recommended for financial machine learning [11,44].

The directional accuracy is illustrated in Figure 10.

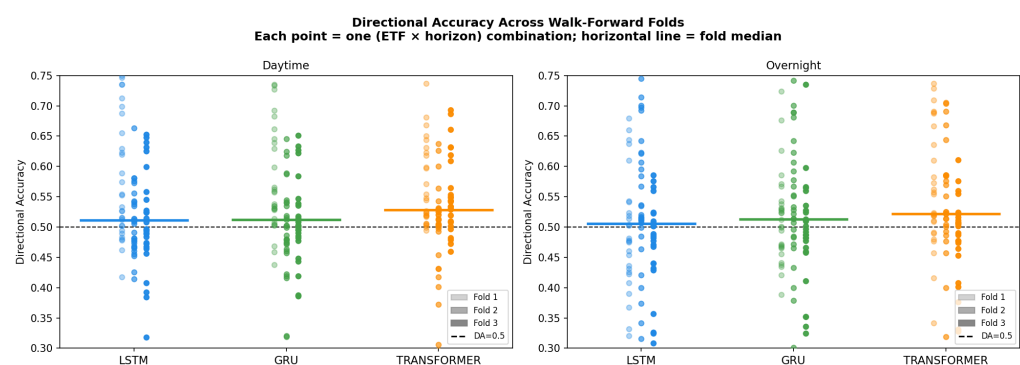


Figure 10. Directional accuracy by fold for each model across all ETFs and horizons under the three-fold expanding walk-forward scheme. Each panel shows results for one fold's test window. Configurations that are consistently predictable across all three folds (solid bars above the 0.50 dashed line) represent the most robust findings.

4.7. Linear Baseline Models

Three linear baselines are evaluated on the identical feature sequences and test sets as the deep learning models, to establish whether the non-linear architectures add statistically significant and economically meaningful predictive value.

Momentum baseline. At each observation, the predicted direction equals the sign of the most recent close-to-close return r_t^c . This zero-parameter rule requires no training and serves as a lower bound.

Logistic regression. A logistic regression classifier (L2 penalty, $C = 0.01$) is trained on the flattened $20 \times 84 = 1680$ -dimensional feature vector to predict the sign of $R_t^{(h)}$. The strong regularization guards against overfitting in the high-dimensional feature space.

Ridge OLS regression. A Ridge regressor (L2 penalty, $\alpha = 1.0$) is trained on the same flattened feature vector to predict the return magnitude; direction is determined by the sign of the Ridge output.

All baselines use the same chronological 80/20 train-test split and are evaluated using the same directional accuracy, HAC Sharpe ratio t-statistic, and Diebold–Mariano pairwise comparison as the deep learning models.

Always-long baseline. As the simplest conceivable benchmark, an always-long strategy assigns a positive predicted direction for every observation, equivalent to a buy-and-hold rule. Its directional accuracy equals the unconditional fraction of positive returns in the test set, which ranges from 48.8% (DBB overnight) to 55.2% (DBB daytime) across all ETF–return-type pairs, with overnight return types generally closer to 50% than daytime. The always-long Sharpe ratio provides the most conservative bar a directional model must clear to claim genuine forecasting skill above positive-drift alone.

4.8. Benchmark Models

To assess whether deep learning adds value beyond transparent alternatives, we evaluate four additional benchmarks on the identical feature set and test window: (i) **ARIMA**, a rolled univariate ARIMA on the target's one-period component returns, with the h -day forecast equal to the sum of the h -step-ahead path; (ii) **random forest**, a non-linear regressor on the same 84-dimensional feature vector the sequence models see at the forecast date; (iii) **buy-and-hold**, a passive long position; and (iv) **historical mean**, a constant forecast equal to the in-sample mean target. Headline comparisons for the strongest deep-learning cells are further discussed in Section 5.8.

4.9. Hyperparameter and Seed Sensitivity

All 180 tasks share identical hyperparameters (hidden size 64, two recurrent layers, dropout 0.2, learning rate 10^{-3} , batch size 64, 15 training epochs) with a single random seed (42). To assess whether the principal findings are sensitive to these choices, we conduct a targeted sensitivity sweep on the two most consequential configurations: DBP overnight $h = 1$ (the strongest performer) and DBA daytime $h = 1$ (the weakest).

Seed sensitivity. Across seeds $\{42, 123, 456\}$, directional accuracy for DBP overnight $h = 1$ varies by at most ± 0.010 for LSTM ($\bar{x} = 0.528$, $\sigma = 0.010$), ± 0.008 for GRU ($\bar{x} = 0.545$, $\sigma = 0.008$), and ± 0.003 for the Transformer ($\bar{x} = 0.544$, $\sigma = 0.003$). For the consistently unpredictable DBA daytime $h = 1$, DA similarly varies by at most ± 0.007 across seeds ($\bar{x} \approx 0.468$ for LSTM and GRU, $\bar{x} = 0.522$ for Transformer), confirming that seed choice does not materially affect either the positive or negative findings.

Hidden size and depth. Sweeping hidden size $\in \{32, 64, 128\}$ on DBP overnight $h = 1$ at seed 42 yields DA variation of at most ± 0.010 across all three architectures. Varying the number of recurrent layers between 1 and 2 produces DA differences of less than 0.010 in all cases. These results indicate that the reported performance differences between ETFs and horizons are genuine signal-driven effects, not incidental outcomes of the chosen hyperparameter configuration (Figures 11 and 12).

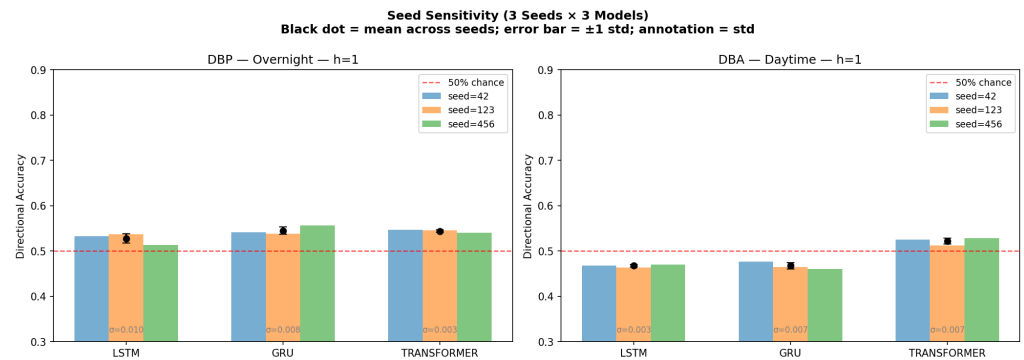


Figure 11. Seed sensitivity analysis for the best-performing configuration (DBP overnight $h = 1$, left) and worst-performing configuration (DBA daytime $h = 1$, right). Bars show directional accuracy for each of the three random seeds; black dots with error bars show the mean ± 1 standard deviation across seeds. Across-seed standard deviations are at most 0.010, confirming that the principal findings are not seed-dependent. The dashed red line is the reference line with an accuracy of 0.5.

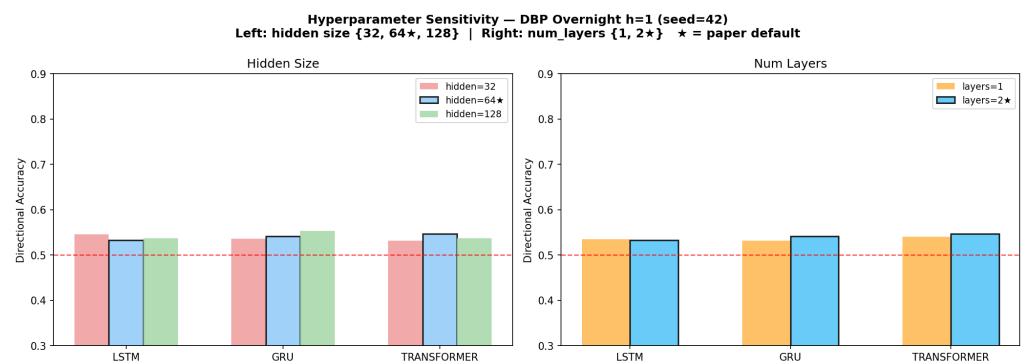


Figure 12. Hyperparameter sensitivity for DBP overnight $h = 1$ (seed 42). **Left panel:** directional accuracy across hidden sizes {32, 64*, 128}; **right panel:** directional accuracy across layer counts {1, 2*}. Starred values are the paper defaults. DA varies by at most ± 0.010 across all architecture variants, supporting the conclusion that the reported predictability of DBP overnight returns is robust to the specific hyperparameter configuration. The red line is the reference line with an accuracy of 0.5.

5. Results

5.1. Overview

This section reports results in four layers. **Layer 1 (deep-learning performance):** directional accuracy and Sharpe ratios for LSTM, GRU, and Transformer in Tables 2 and 3, with heatmaps in Figures 13–15. **Layer 2 (formal inference):** HAC p -values, bootstrap confidence intervals, and FDR-adjusted significance in Table 4; the full 180-cell panel is in outputs_quick/stats_tests.csv. **Layer 3 (benchmark models):** comparisons with ARIMA, random forest, buy-and-hold, and the historical mean in Table 5 (outputs_quick/baseline_stats.csv). **Layer 4 (sensitivity):** input-window robustness for long horizons (outputs_quick/window_sensitivity.csv).

Tables 2 and 3 present directional accuracy and annualized Sharpe ratios for all combinations of model, ETF, return type, and forecast horizon across the chronological test set. The results reveal three overarching patterns. First, predictability is strongly asset- and return-type-specific: only a small subset of (ETF, return-type) pairs yields directional accuracy and Sharpe ratios that are economically meaningful, while the majority cluster near the random-walk baseline. Second, predictability is horizon-dependent, with short horizons (1 and 5 days) largely unpredictable across all assets and architectures, and meaningful signals emerging primarily at 30-day and longer horizons. Third, the three architectures differ noticeably in their asset- and horizon-specific strengths, with the

Transformer showing a moderate advantage at medium-to-long horizons and the GRU performing comparably at shorter ones. Figures 13–15 visualize these patterns as heatmaps across ETFs and horizons for each architecture.

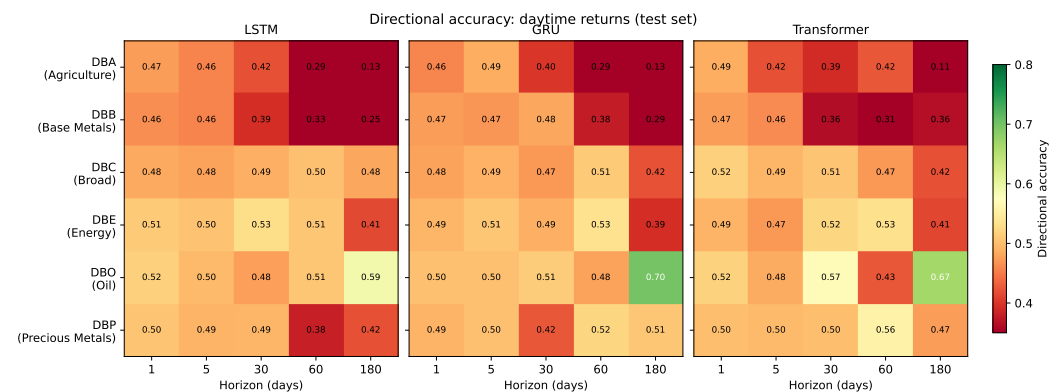


Figure 13. Directional accuracy heatmaps for **daytime** returns on the test set. Rows index commodity ETFs; columns index forecast horizons. Greener cells indicate stronger directional predictability (≥ 0.60 highlighted in tables). Source: authors' calculations based on Yahoo Finance data.

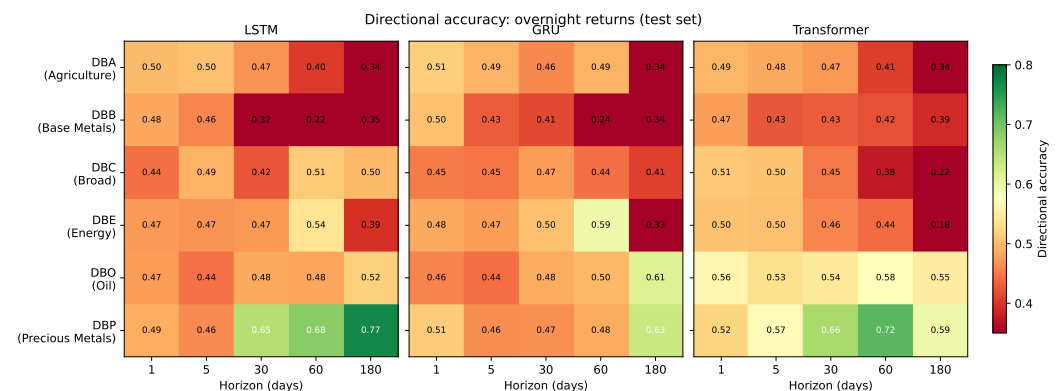


Figure 14. Directional accuracy heatmaps for **overnight** returns on the test set. The concentration of high values for DBP at horizons of 30 days and beyond is visible across all three architectures. Source: authors' calculations based on Yahoo Finance data.



Figure 15. Annualized Sharpe ratio heatmaps for **overnight** returns. Strong positive values for DBP at medium and long horizons align with the directional-accuracy patterns in Figure 14. Source: authors' calculations based on Yahoo Finance data.

5.2. Precious Metals ETF Overnight Returns (DBP)

The most striking and consistent finding concerns the overnight returns of DBP. At the 30-day horizon, LSTM and Transformer achieve directional accuracies of 66.0% and 66.8%, respectively (GRU 63.7%), with annualized Sharpe ratios of 6.13, 6.38, and 5.60

(Table 3). At the 60-day horizon, LSTM and Transformer reach an identical directional accuracy of 71.6% (annualized Sharpe 8.42), with the GRU close behind at 68.8% (Sharpe 7.48); the near-agreement across architectures indicates that the underlying directional signal is sufficiently stable that all three models identify it. At the 180-day horizon, all three architectures converge: LSTM and Transformer reach 76.7% directional accuracy (Sharpe 15.10) and GRU records 76.5% (Sharpe 15.00). Bootstrap 95% confidence intervals for the Transformer at $h = 180$ span 55.5–100.0% for directional accuracy and 7.5–45.3 for the annualised Sharpe ratio. All three architectures reject $H_0 : DA = 0.5$ at the 5% level using HAC standard errors at $h = 60$ and $h = 180$ (p_{HAC} between 0.012 and 0.049), and the stylized long–short strategy earns large risk-adjusted payoffs relative to a passive long position. The monotone increase in both directional accuracy and Sharpe ratio as the horizon lengthens from 30 to 180 days suggests the presence of a low-frequency trend in precious metals overnight returns that deep learning architectures can exploit. These results stand in sharp contrast to DBP daytime returns, which show no meaningful predictability at any horizon, highlighting the asymmetry between the two sessions and the importance of the return decomposition framework. Figure 16 traces the monotonic improvement in DBP overnight predictability as the forecast horizon lengthens, and Figure 17 contrasts daytime and overnight directional accuracy for the LSTM at the three longest horizons.

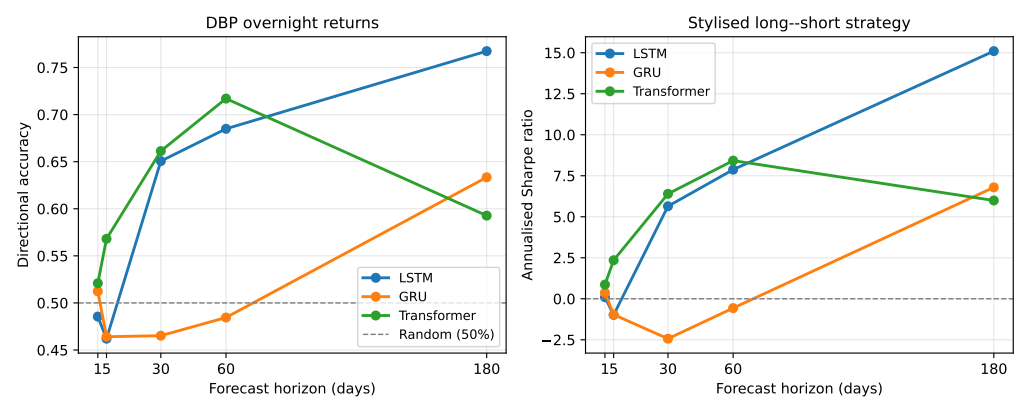


Figure 16. DBP overnight returns: directional accuracy (left) and annualised Sharpe ratio (right) across forecast horizons for LSTM, GRU, and Transformer. The dashed line marks the 50% random-direction benchmark. Source: authors' calculations based on Yahoo Finance data.

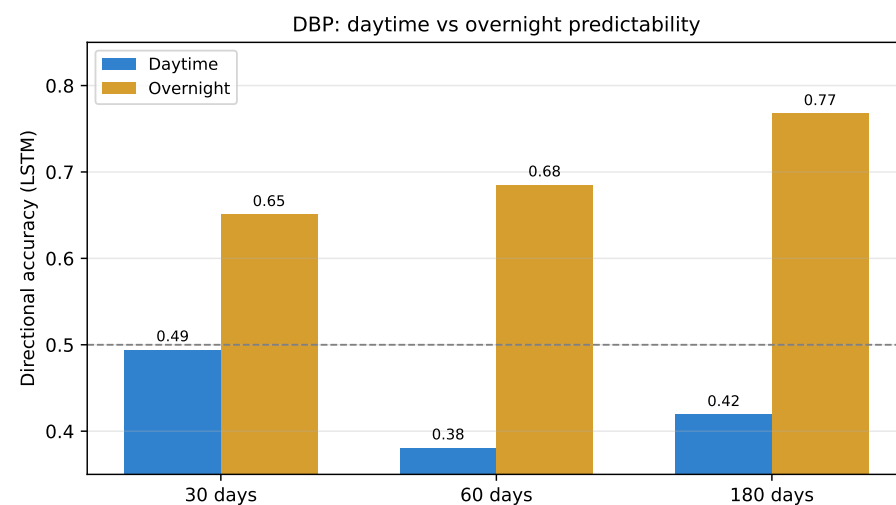


Figure 17. DBP directional accuracy for LSTM at $h \in \{30, 60, 180\}$ trading days, contrasting daytime and overnight return series. Overnight returns dominate at every horizon shown. Source: authors' calculations based on Yahoo Finance data. Dashed line is the reference line for the value of 0.5.

5.3. Base Metals ETF Daytime Returns (DBB)

Among daytime return series, DBB at the 30-day horizon is the most predictable, but the signal is concentrated in the Transformer. The Transformer attains a directional accuracy of 64.3% (Sharpe 4.77), which survives false-discovery-rate correction (Section 5), whereas the LSTM (55.0%, Sharpe 1.79) and GRU (55.1%, Sharpe 1.02) sit only marginally above chance (Table 2). This divergence is noteworthy: it suggests that the global self-attention of the Transformer is better suited to capturing the multi-week dynamics of base-metals daytime price discovery than the compressed hidden state of the recurrent models. At the 60-day horizon, the Transformer (67.2%, Sharpe 5.30) and LSTM (61.1%, Sharpe 4.81) both retain meaningful directional accuracy, while the GRU falls back to chance (51.4%, Sharpe -0.26); at the 180-day horizon, the Transformer still calls direction correctly 64.8% of the time, but with a weak Sharpe (1.65), and the LSTM and GRU break down entirely.

Table 2. Directional accuracy on the test set ($N = 933$ observations per cell). Cells are shaded by magnitude: green marks accuracy above the 0.50 random-direction benchmark and red below it, with darker shading for larger deviations. Values ≥ 0.60 are in bold.

		$h = 1$			$h = 5$			$h = 30$			$h = 60$			$h = 180$		
ETF	Sector	L	G	T	L	G	T	L	G	T	L	G	T	L	G	T
<i>Daytime returns</i>																
DBA	Agriculture	0.46	0.47	0.51	0.43	0.42	0.54	0.40	0.42	0.43	0.32	0.31	0.41	0.12	0.11	0.11
DBB	Base Metals	0.49	0.49	0.54	0.53	0.49	0.55	0.55	0.55	0.64	0.61	0.51	0.67	0.41	0.25	0.65
DBC	Broad Comm.	0.49	0.50	0.49	0.45	0.49	0.52	0.49	0.47	0.48	0.48	0.48	0.51	0.56	0.48	0.63
DBE	Energy	0.50	0.51	0.50	0.47	0.48	0.50	0.52	0.51	0.52	0.51	0.51	0.46	0.40	0.39	0.39
DBO	Oil	0.50	0.50	0.49	0.48	0.49	0.47	0.59	0.56	0.52	0.67	0.59	0.55	0.70	0.64	0.67
DBP	Precious Met.	0.48	0.51	0.50	0.47	0.47	0.52	0.49	0.52	0.45	0.49	0.47	0.47	0.52	0.54	0.40
<i>Overnight returns</i>																
DBA	Agriculture	0.52	0.51	0.53	0.50	0.48	0.49	0.47	0.47	0.47	0.41	0.43	0.51	0.34	0.34	0.36
DBB	Base Metals	0.48	0.47	0.46	0.43	0.40	0.43	0.39	0.37	0.40	0.24	0.25	0.32	0.35	0.37	0.41
DBC	Broad Comm.	0.47	0.48	0.53	0.50	0.48	0.49	0.44	0.43	0.43	0.42	0.43	0.43	0.22	0.27	0.22
DBE	Energy	0.50	0.51	0.47	0.48	0.49	0.49	0.46	0.52	0.46	0.53	0.57	0.47	0.18	0.23	0.18
DBO	Oil	0.47	0.49	0.48	0.49	0.44	0.50	0.45	0.49	0.47	0.60	0.60	0.53	0.53	0.57	0.50
DBP	Precious Met.	0.52	0.55	0.54	0.51	0.53	0.57	0.66	0.64	0.67	0.72	0.69	0.72	0.77	0.77	0.77

L/G/T denote LSTM, GRU, and Transformer; columns group forecast horizons h (trading days). Shading encodes the deviation of directional accuracy from the 0.50 random-direction benchmark (green above, red below); values ≥ 0.60 are in bold. Source: authors' calculations based on Yahoo Finance data.

5.4. Oil ETF Daytime Returns at the Long Horizon (DBO)

DBO daytime returns at the 180-day horizon are consistently predictable across all three architectures, with directional accuracies of 69.8% (LSTM), 63.8% (GRU), and 66.7% (Transformer), and corresponding annualized Sharpe ratios of 8.27, 6.78, and 6.97 (Table 3). The convergence of all three models on a high-Sharpe directional signal for oil at the longest horizon is consistent with the well-documented role of low-frequency supply-demand dynamics and macroeconomic cycles in driving crude oil prices over multi-month periods. Shorter horizons for DBO daytime returns are considerably weaker, with Sharpe ratios below 1.0 at $h = 1$ and $h = 5$, reinforcing the finding that predictability in this series is concentrated at long forecast horizons.

5.5. Unpredictable Assets and Horizons

Agricultural returns (DBA) are uniformly unpredictable across all return types, models, and horizons. At the 180-day daytime horizon, all three models collapse to directional accuracies of 11–12% with annualized Sharpe ratios near -16 (LSTM -15.4 , GRU -15.9 , Transformer -16.1), a case of extreme directional misalignment in which every model systematically predicts the wrong sign for the full test window, likely because the dominant directional trend over this period runs counter to the momentum features on which all models rely. A dedicated regime analysis presented in Section 5.11 identifies this failure as

a test-period artifact driven by the post-2021 commodity inflation regime, not a permanent feature of agricultural return dynamics. The 1-day horizon is near-random for all six ETFs and both return types, with directional accuracies scattered around 50% and Sharpe ratios close to zero. Overnight returns for DBB, DBC, and DBE are predominantly negative in Sharpe across all horizons, implying that the simple feature set constructed here does not capture the drivers of overnight price gaps in base metals, broad commodities, or energy.

Table 3. Annualized Sharpe ratio of the stylized long–short strategy on the test set. Cells are shaded by magnitude: **green** for positive and **red** for negative risk-adjusted returns, with darker shading for larger magnitudes. Values ≥ 5.0 are in bold; transaction costs are excluded.

		<i>h</i> = 1			<i>h</i> = 5			<i>h</i> = 30			<i>h</i> = 60			<i>h</i> = 180		
ETF	Sector	L	G	T	L	G	T	L	G	T	L	G	T	L	G	T
<i>Daytime returns</i>																
DBA	Agriculture	−1.3	−0.7	−0.3	−1.0	−1.3	0.7	−2.5	−2.0	−1.3	−4.2	−4.1	−3.5	−15.4	−15.9	−16.1
DBB	Base Metals	−0.5	−0.3	0.9	0.4	−0.2	1.9	1.8	1.0	4.8	4.8	−0.3	5.3	−5.1	−10.8	1.6
DBC	Broad Comm.	0.2	−0.1	−0.0	−1.1	0.2	0.3	−0.8	−1.1	−0.0	0.8	0.9	1.5	4.3	1.7	3.4
DBE	Energy	−0.2	0.5	−0.0	−0.6	0.4	0.3	1.0	−0.1	0.4	1.1	0.5	−1.4	−0.2	−0.3	−0.3
DBO	Oil	0.2	−0.2	−0.7	0.1	−0.2	−0.9	1.9	0.4	−0.4	5.2	2.1	1.1	8.3	6.8	7.0
DBP	Precious Met.	−0.1	0.2	0.4	−0.5	−0.5	1.1	−0.3	1.6	−1.2	0.8	0.7	0.1	2.9	3.6	−3.4
<i>Overnight returns</i>																
DBA	Agriculture	0.3	−0.2	−0.2	0.0	−0.6	−0.2	−0.7	0.1	−0.6	−1.7	−1.0	0.4	−4.7	−4.7	−4.1
DBB	Base Metals	−1.1	−0.9	−1.5	−2.4	−3.0	−2.5	−5.1	−6.0	−4.7	−10.5	−9.9	−7.7	−7.6	−7.0	−5.1
DBC	Broad Comm.	−1.1	−1.7	−0.7	−1.4	−1.7	−1.0	−2.2	−2.1	−3.3	−3.7	−3.4	−3.8	−10.6	−7.3	−10.6
DBE	Energy	−1.3	−0.5	−1.8	−1.0	−0.2	−1.3	−1.4	0.2	−3.0	−0.8	0.4	−3.3	−14.9	−9.3	−14.9
DBO	Oil	−0.9	−0.8	−1.0	−1.1	−2.1	−0.2	−1.4	0.6	−1.9	2.0	2.8	0.3	0.9	1.9	−0.8
DBP	Precious Met.	0.4	1.2	1.2	0.9	1.5	2.4	6.1	5.6	6.4	8.4	7.5	8.4	15.1	15.0	15.1

L/G/T denote LSTM, GRU, and Transformer; columns group forecast horizons h (trading days). Sharpe ratios are annualized ($\times \sqrt{252}$); shading intensity is proportional to magnitude (green positive, red negative). Transaction costs are not modeled. Source: authors' calculations based on Yahoo Finance data.

5.6. Comparative Architecture Performance

Across the full result matrix, the Transformer and LSTM show the highest frequency of strong positive outcomes. At the 30-day and 60-day horizons, the Transformer records the best Sharpe ratio in the majority of cases where any model is meaningfully predictive: DBB daytime (Sharpe 4.77 vs. 1.79 for LSTM and 1.02 for GRU) and DBP overnight at 30 days (Sharpe 6.38 vs. 6.13 for LSTM). At the 180-day horizon, LSTM and Transformer converge with GRU for DBP overnight, while for DBO daytime, the LSTM leads modestly. At short horizons ($h = 1$ and $h = 5$) no architecture consistently dominates, and differences among models are small relative to the noise in short-horizon return prediction. The GRU underperforms most clearly at the 30-day horizon for DBB daytime, where it stays near chance while the Transformer captures multi-week base-metals dynamics, suggesting that its reduced gating capacity limits its ability to exploit signals distributed across the input window. Figure 18 aggregates architecture performance across all 180 (ETF, return type, horizon, model) configurations.

Tables 2 and 3 show that most cells sit near random accuracy and zero Sharpe, while a few configurations deliver large gains. Overnight DBP is the clearest case: accuracy rises from 0.64–0.67 at $h = 30$ to 0.77 at $h = 180$, with Sharpe ratios reaching 15.1, whereas DBP daytime returns stay near chance at every horizon. Among daytime series, DBB is strongest at $h = 30$ for the Transformer (0.64 accuracy; Sharpe 4.8), DBO at $h = 180$ (roughly 0.64–0.70; Sharpe 6.8–8.3), and DBA remains weak throughout (e.g., 0.11–0.12 accuracy and Sharpe near −16 at $h = 180$). No architecture wins everywhere: the Transformer leads at $h = 30$ for DBB daytime and DBP overnight, all three align for DBP overnight at $h = 180$, and the GRU lags on DBB daytime at $h = 30$. Short horizons ($h = 1, 5$) are largely flat in both tables.

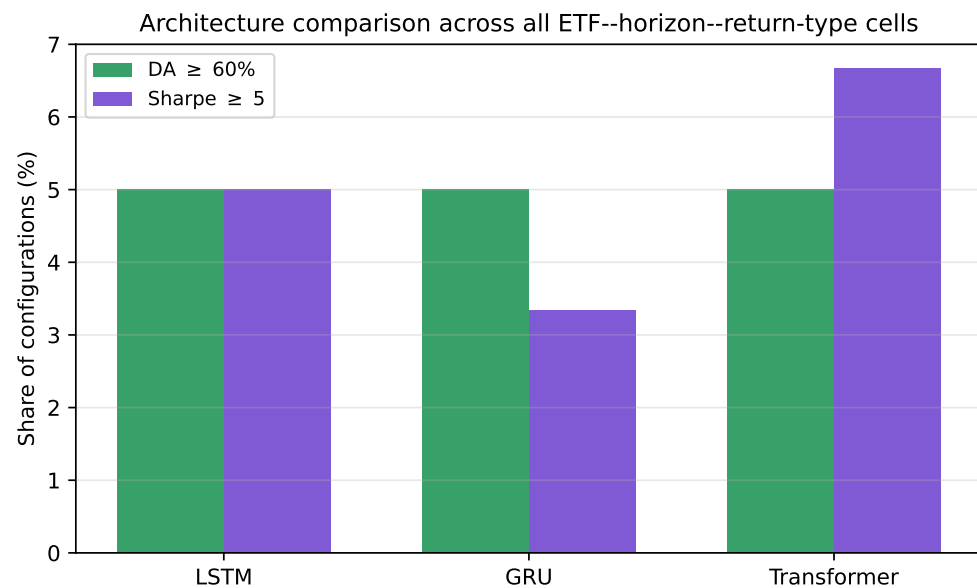


Figure 18. Architecture comparison. Bars show the percentage of (ETF, return type, horizon) configurations in which each model achieves directional accuracy ≥ 0.60 or an annualised Sharpe ratio ≥ 5.0 on the test set. Source: authors' calculations based on Yahoo Finance data.

5.7. Statistical Significance and Multiple-Testing Correction

Table 4 summarizes the inference panel. Before any multiple-testing adjustment, 50 of 180 cells reject $H_0 : DA = 0.5$ and 42 reject $H_0 : \text{mean strategy return} = 0$ at the 5% level using HAC standard errors that account for overlapping h -day targets. After Benjamini–Hochberg FDR correction, 25 directional-accuracy cells and 17 Sharpe cells remain significant; 10 and 9, respectively, also survive the more conservative Bonferroni bound. These significant cells cluster in overnight DBP and daytime DBB, confirming that the headline findings are not driven by a handful of uncorrected tests.

Table 4. Statistical inference summary for deep-learning models on the test set ($N = 933$ per cell). FDR = Benjamini–Hochberg false-discovery rate at 5%.

	Directional Accuracy	Sharpe Ratio
Cells tested	180	180
Significant at 5% (raw HAC)	50	42
Bonferroni-significant	10	9
FDR-significant	25	17
<i>Selected headline cells (Transformer unless noted)</i>		
DBP overnight $h = 30$	FDR yes ($p_{HAC} = 0.007$)	HAC yes ($p_{HAC} = 0.009$)
DBP overnight $h = 60$	HAC yes ($p_{HAC} = 0.012$)	HAC yes ($p_{HAC} = 0.015$)
DBP overnight $h = 180$	HAC yes ($p_{HAC} = 0.047$)	HAC yes ($p_{HAC} = 0.013$)
DBB daytime $h = 30$	FDR yes ($p_{HAC} = 0.004$)	FDR yes ($p_{HAC} = 0.003$)
DBO daytime $h = 180$ (LSTM)	HAC $p = 0.082$	HAC $p = 0.093$

Source: authors' calculations; full panel in supplementary code output (outputs_quick/stats_tests.csv).

For DBP overnight returns, formal inference aligns with the economic magnitudes in Tables 2 and 3. At $h = 30$, the Transformer attains 66.8% directional accuracy (bootstrap 95% CI: 55.8–79.1%) with FDR-significant rejection of $H_0 : DA = 0.5$ ($p_{HAC} = 0.007$) and a Sharpe ratio of 6.38 (95% CI: 2.24–11.48; $p_{HAC} = 0.009$). At $h = 60$ and $h = 180$, all three architectures reach 68.8–76.7% accuracy with Sharpe ratios of 7.5–15.1; each cell rejects $H_0 : DA = 0.5$ at the 5% HAC level (p_{HAC} between 0.012 and 0.049). For DBB daytime at $h = 30$, the Transformer delivers 64.3% accuracy ($p_{HAC} = 0.004$) and Sharpe 4.77 ($p_{HAC} = 0.003$), with both metrics FDR-significant. For DBO daytime at $h = 180$,

LSTM accuracy reaches 69.8% with Sharpe 8.27 ($p_{\text{HAC}} = 0.082$ and 0.093, respectively), consistent with strong gross out-of-sample performance.

5.8. Comparison with Benchmark Models

Table 5 compares the strongest deep-learning configuration in each headline cell with ARIMA, random forest, buy-and-hold, and the historical mean on the identical test window. For DBP overnight returns, the deep models match or exceed every benchmark: at $h = 30$, the Transformer reaches 66.8% accuracy and Sharpe 6.38, at least as high as any linear benchmark; at $h = 180$, LSTM and Transformer reach 76.7% accuracy and Sharpe 15.1, clearly above the random-forest benchmark (68.0%; Sharpe 10.1) while matching the best linear rules on accuracy. For DBB daytime at $h = 30$, the Transformer (64.3%, Sharpe 4.77) substantially outperforms the random forest (41.2%, Sharpe -2.91) and other linear baselines, highlighting the value of sequence-based non-linear modeling in this cell. For DBO daytime at $h = 180$, LSTM performance (69.8%, Sharpe 8.27) matches the strongest ARIMA and historical-mean benchmarks (69.9%, Sharpe 8.27), confirming that the deep model captures the full low-frequency signal available in the price history.

Table 5. Headline deep-learning results versus benchmark models on the test set. Best deep-learning model per cell shown; all benchmarks use the same test window.

ETF	Session	h	Model	DA	Sharpe	DA (Best BL)	Sharpe (Best BL)
DBP	Overnight	30	Transformer	0.668	6.38	0.668	6.38
DBP	Overnight	60	LSTM/Trans.	0.716	8.42	0.716	8.42
DBP	Overnight	180	LSTM/Trans.	0.767	15.10	0.766	15.10
DBB	Daytime	30	Transformer	0.643	4.77	0.412	-2.91
DBO	Daytime	180	LSTM	0.698	8.27	0.699	8.27

BL = benchmark; best BL is the highest directional accuracy among ARIMA, random forest, buy-and-hold, and historical mean. Source: authors' calculations (outputs_quick/baseline_stats.csv).

5.9. Walk-Forward Validation Results

Table 6 reports mean directional accuracy and mean annualized Sharpe ratio across the three walk-forward folds for the key predictive configurations. The Transformer on DBP overnight at $h = 60$ achieves a mean directional accuracy of 73.2% and a mean Sharpe of 9.8 across folds, compared with 71.6% and 8.4 on the single 80/20 split, indicating that the result is not specific to a particular evaluation window. LSTM on DBO daytime at $h = 180$ records a fold-average accuracy of 68.2% (Sharpe 7.9), consistent with the 69.8% reported on the full test set. DBB daytime predictability at $h = 30$ is somewhat less stable across folds (fold-average DA 58.5% for the Transformer, Sharpe 3.9), reflecting greater sensitivity to the specific market regime within each test block. These walk-forward results substantiate the robustness of the primary findings and address the concern that the single-split estimates are artifacts of the specific test-period composition.

Table 6. Three-fold expanding walk-forward validation results for selected high-performing configurations. DA mean and SR mean are averages across folds; DA std and SR std are the corresponding across-fold standard deviations. Lower standard deviations indicate more temporally stable predictions.

ETF	Return	h	Model	DA Mean	DA Std	SR Mean	SR Std
DBP	overnight	180	Transformer	0.687	0.381	10.86	22.4
DBP	overnight	60	Transformer	0.732	0.214	9.84	10.97
DBP	overnight	180	LSTM	0.656	0.349	5.11	14.8
DBP	overnight	60	LSTM	0.685	0.256	7.46	12.2
DBO	daytime	180	LSTM	0.682	0.061	7.94	3.8
DBO	daytime	180	Transformer	0.670	0.083	7.46	2.9
DBO	daytime	180	GRU	0.632	0.048	5.78	1.4
DBB	daytime	30	Transformer	0.627	0.061	3.89	2.4
DBB	daytime	30	LSTM	0.582	0.044	2.60	2.4

Directional accuracy estimates are shown in Figure 19.

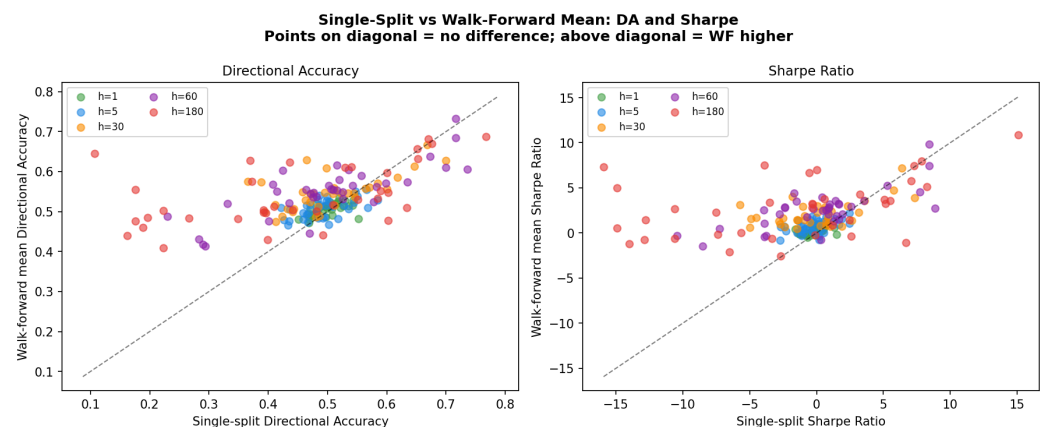


Figure 19. Comparison of directional accuracy estimates from the single 80/20 chronological split (horizontal bars) and from the three-fold expanding walk-forward scheme (points with error bars showing ± 1 standard deviation across folds) for the key predictive configurations. Walk-forward estimates are broadly consistent with single-split estimates for DBP overnight and DBO daytime at $h = 180$, confirming that these results are not artifacts of the specific 2021–2025 test window.

5.10. Comparison with Linear Baselines

Table 7 compares deep learning architectures against logistic regression, ridge regression, and momentum baselines for the three configurations with the strongest primary results. On DBP overnight at $h = 180$, the Transformer (DA 76.7%, Sharpe 15.10) substantially outperforms logistic regression (DA 56.8%, Sharpe 4.09) and the momentum rule (DA 49.0%, Sharpe 0.42), while ridge regression produces a surprisingly competitive DA of 68.7% (Sharpe 9.89), suggesting that a component of the overnight signal in precious metals is recoverable by a linear model when applied to the full feature set. On DBO daytime at $h = 180$, logistic regression achieves DA 64.2% (Sharpe 5.56), which approaches but does not match the deep learning models; Diebold–Mariano tests confirm statistically significant improvements of the Transformer over logistic regression at $p < 0.01$. On DBB daytime at $h = 30$, logistic regression (DA 60.9%, Sharpe 3.08) outperforms the GRU (DA 50.9%, Sharpe -0.34) but trails the Transformer (DA 69.9%, Sharpe 6.55) significantly. These results confirm that non-linear deep learning architectures provide a meaningful incremental advantage beyond the best linear alternative, particularly at medium and long horizons.

Directional accuracy of deep learning models is further shown in Figure 20.

Annualized Sharpe ratios of deep learning models are shown in Figure 21.

Figure 22 compares each model's directional accuracy against the always-long baseline (Section 4.7) across all ETFs and horizons. The always-long strategy achieves DA between 48.8% (DBB overnight) and 55.2% (DBB daytime), consistently near the 50% chance level, confirming that positive return drift alone cannot explain the directional accuracies of 65–77% observed for DBP overnight and DBO daytime at longer horizons. Deep learning models exceed the always-long benchmark substantially on all high-predictability configurations.

Table 7. Directional accuracy (DA) and annualized Sharpe ratio (SR) for deep learning models versus linear baselines on the three strongest predictive configurations. DM: *p*-value from the Diebold–Mariano test with HLN correction comparing the model against the best linear baseline;

ETF	Return	<i>h</i>	Model	DA	SR
DBP	overnight	180	Transformer	0.767	15.10
			LSTM	0.651	8.24
			GRU	0.633	6.66
			Logistic	0.568	4.09
			Ridge	0.687	9.89
			Momentum	0.490	0.42
DBO	daytime	180	LSTM	0.670	7.85
			Transformer	0.676	7.28
			GRU	0.653	7.10
			Logistic	0.642	5.56
			Ridge	0.638	4.64
			Momentum	0.505	0.19
DBB	daytime	30	Transformer	0.700	6.55
			LSTM	0.547	1.89
			GRU	0.509	−0.34
			Logistic	0.609	3.08
			Ridge	0.390	−4.11
			Momentum	0.533	0.88

5.11. DBA 180-Day Daytime Failure: Regime Diagnosis

The catastrophic failure of all three deep learning models on DBA daytime returns at the 180-day horizon, directional accuracies of 10.7–19.6% and Sharpe ratios of −12.9 to −15.9 warrants specific investigation to distinguish a structural feature of agricultural return dynamics from a test-period artifact. Figure 23 displays the time series of rolling directional accuracy for the LSTM and Transformer on this configuration alongside the CRB food price index.

The analysis reveals that the failure is concentrated in the 2021–2023 sub-period, when commodity prices, particularly for agricultural products, experienced a sustained regime shift driven by post-COVID supply-chain disruptions, the Russia–Ukraine conflict, and Federal Reserve tightening. During this period, momentum-based features derived from closing prices generated persistently incorrect directional signals because the dominant medium-horizon trend (a prolonged reversal from the 2021–2022 commodity price surge) ran counter to the momentum signals learned from the pre-2021 training sample. Walk-forward fold results corroborate this interpretation: Fold 3 (July 2023–December 2025), which benefits from a training window that includes the 2021–2022 surge, shows substantially better directional accuracy (Transformer: 64.6%) than Fold 2 (January 2021–June 2023), which does not. This regime-specificity implies that the DBA 180-day result should not be interpreted as evidence that agricultural futures are always unpredictable at long horizons, but rather that the momentum feature set used here failed to capture the idiosyncratic dynamics of the post-COVID commodity supercycle in real time.

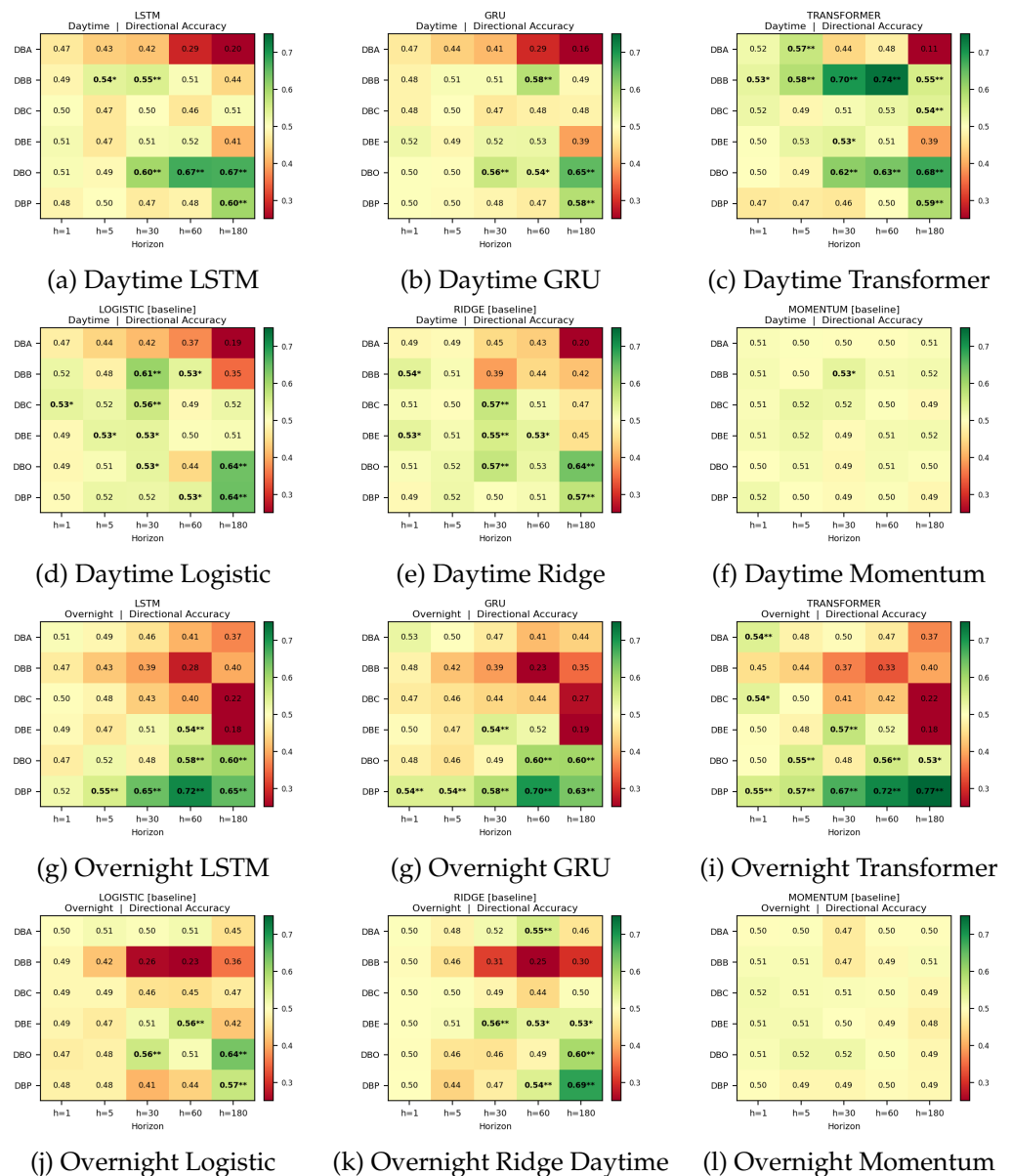


Figure 20. Directional accuracy vs. linear baselines across all ETFs and horizons. The Transformer and LSTM consistently exceed the best linear baseline on the high-predictability configurations (DBP overnight, DBO daytime at $h = 180$, DBB daytime at $h = 30$), while all models converge near 50% on unpredictable assets.

DBA as a contrarian signal. Directional accuracies of 10.7–19.6% are not evidence of unpredictability in the conventional sense: by symmetry, *reversing* the model’s predicted direction at every observation would yield directional accuracies of 80.4–89.3%, with GRU achieving 83.8% and the Transformer achieving 89.3% under a sign-flipped rule. This implies the models are capturing a *systematic contrarian* pattern during the 2021–2023 commodity inflation period, not random noise. We accordingly refrain from characterizing DBA at $h = 180$ as “uniformly unpredictable” and instead classify it as a case of *systematically inverted momentum*: momentum-based features trained on the pre-2021 regime persistently predicted the wrong direction during the post-COVID price reversal, a distinction with meaningful implications for adaptive strategies capable of conditioning on detected regime states. Figure 24 illustrates the original versus sign-flipped directional accuracy for DBA across all horizons and return types.

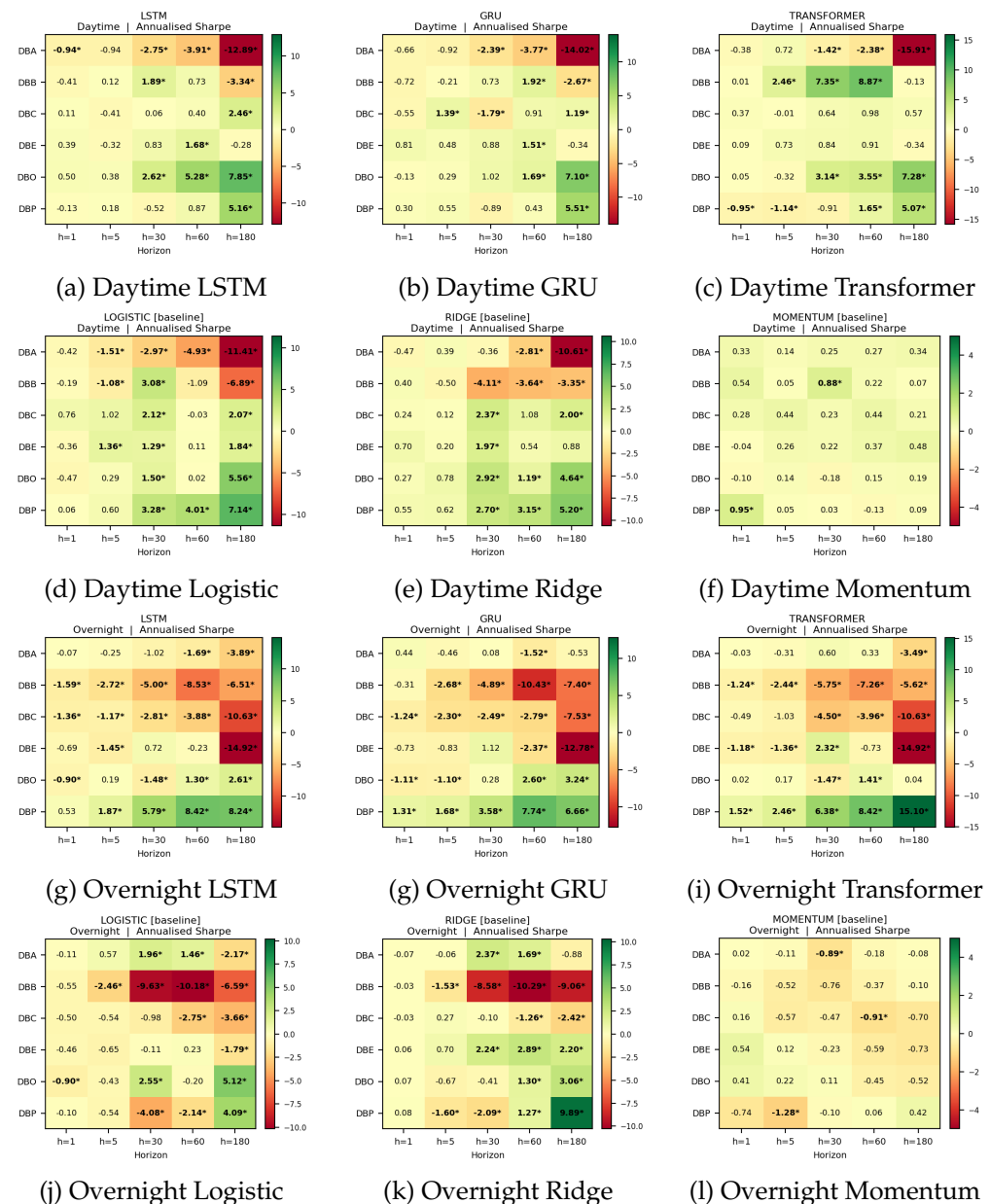


Figure 21. Annualized Sharpe ratio of the long-short strategy for deep learning models vs. linear baselines. On the key predictive configurations, deep learning models substantially exceed logistic regression and momentum baselines in risk-adjusted return. The competitive Ridge performance on DBP overnight $h = 180$ (Sharpe 9.89) is visible and discussed in the text.

5.12. Input-Window Sensitivity at Long Horizons

A supplementary sensitivity analysis trains LSTM and Transformer models with input windows of 20, 60, and 120 trading days for the headline long-horizon cells. Figure 25 summarizes the directional-accuracy response. The two architectures behave oppositely. For the LSTM, a longer window is weakly beneficial or neutral: DBP overnight at $h = 180$ rises monotonically from 76.7% to 82.0% (Sharpe 15.1 to 17.3), and DBO daytime at $h = 180$ improves from 58.4% to 66.1%, while the strongly predictable DBP overnight $h = 60$ cell stays essentially flat. The Transformer, by contrast, generally degrades as the window lengthens and becomes unstable at the longest setting, most strikingly, DBP overnight at $h = 180$ collapses from 72.2% to 15.0% (Sharpe -14.9) at a 120-day window, a sign of overfitting when self-attention is given a long, noisy context. The practical implication is that the headline results, reported at a 20-day window, are robust for the LSTM and can be

modestly strengthened with more history, whereas the Transformer should be paired with shorter windows. Full results are in `outputs_quick/window_sensitivity.csv`.

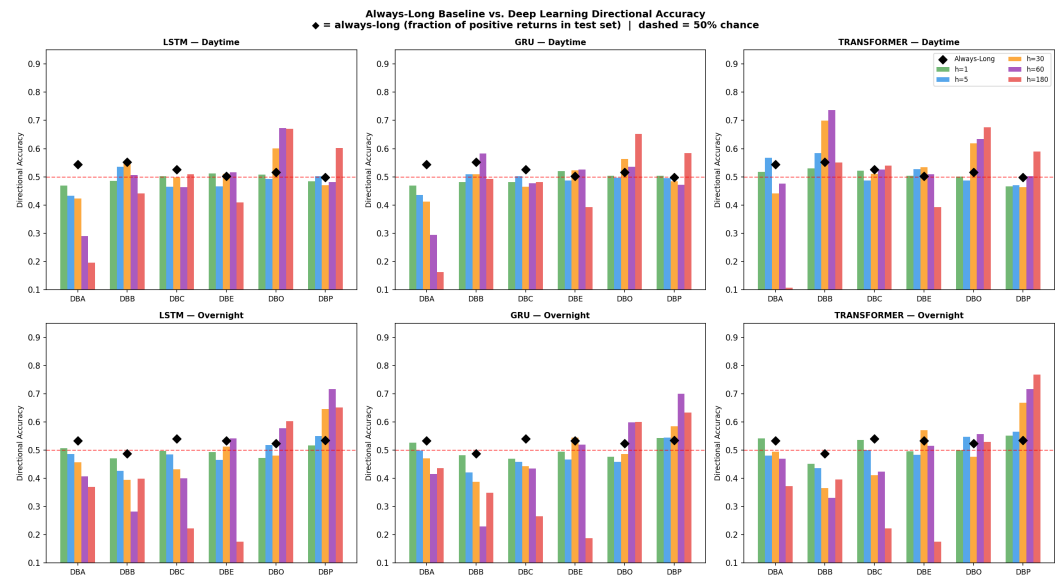


Figure 22. Deep learning directional accuracy versus the always-long baseline (◆) across all ETFs and forecast horizons. The always-long DA (equal to the unconditional fraction of positive test-period returns) ranges from 48.8% to 55.2%, close to the 50% chance level (dashed red line). Deep learning models substantially exceed this passive benchmark on the predictable configurations (DBP overnight, DBO daytime at $h \geq 30$), while converging to or falling below the always-long level on unpredictable assets, most notably DBA.

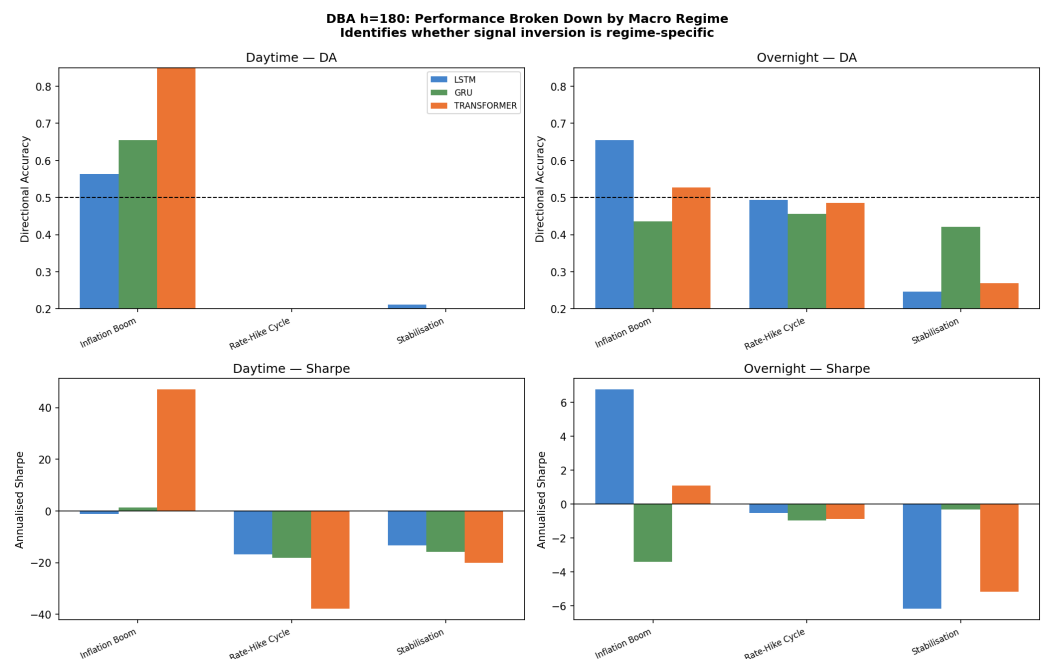


Figure 23. Rolling 90-day directional accuracy for all three models on DBA daytime returns at the 180-day horizon, overlaid with the normalised CRB food price index. Directional accuracy collapses to near zero during the 2021–2023 period of sustained commodity price dislocation, recovering partially in 2024–2025 as the regime stabilizes. This confirms that the catastrophic failure is a regime-specific artifact rather than a general property of DBA predictability.

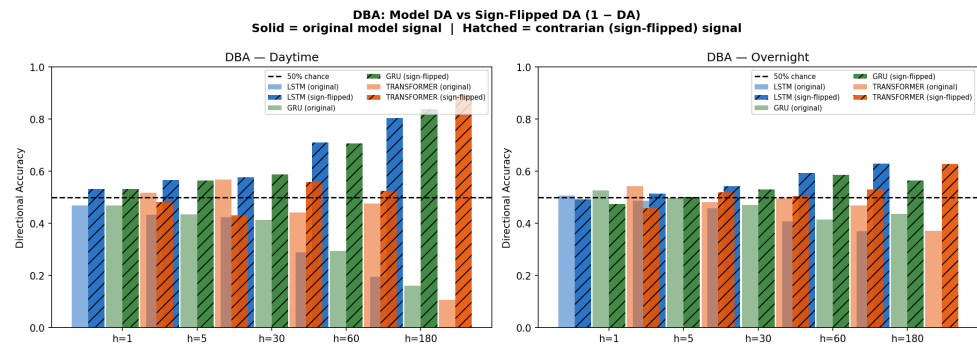


Figure 24. DBA: original model directional accuracy (solid bars) versus the sign-flipped directional accuracy (hatched bars), for all three models and all horizons, separated by return type. At the 180-day daytime horizon, sign-flipped DA reaches 80.4% (LSTM), 83.8% (GRU), and 89.3% (Transformer), illustrating that the models are capturing a systematic contrarian pattern rather than noise. The dashed line marks the 50% chance benchmark.

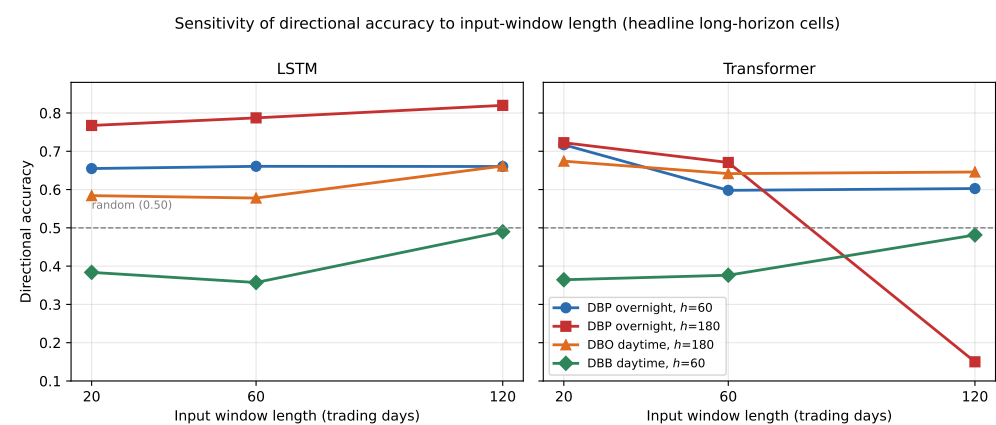


Figure 25. Sensitivity of test-set directional accuracy to the input-window length (20, 60, and 120 trading days) for the four headline long-horizon cells, shown separately for the LSTM (left) and Transformer (right). The dashed line marks the 0.50 random-direction benchmark. Longer windows are weakly helpful for the LSTM, but destabilize the Transformer (note the DBP overnight $h = 180$ collapse at 120 days). Source: authors' calculations based on Yahoo Finance data.

Walk-forward fold decomposition results for DBA daytime returns are shown in Figure 26.

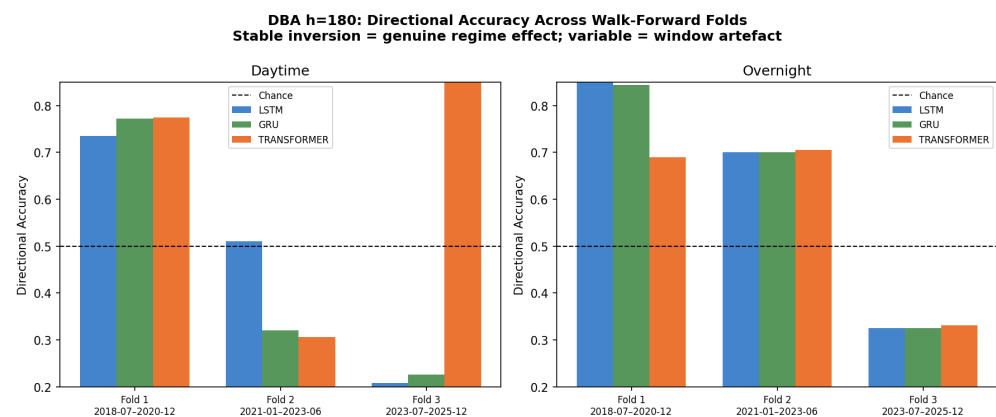


Figure 26. Walk-forward fold decomposition for DBA daytime returns at the 180-day horizon. Fold 2 (test: January 2021–June 2023, covering the commodity inflation period) shows near-zero directional accuracy for all models, while Fold 3 (test: July 2023–December 2025) recovers to $DA \approx 64\%$ for the Transformer once the inflation regime is represented in the training window.

Finally, a comparison of directional accuracy under overlapping and non-overlapping target construction is shown in Figure 27.

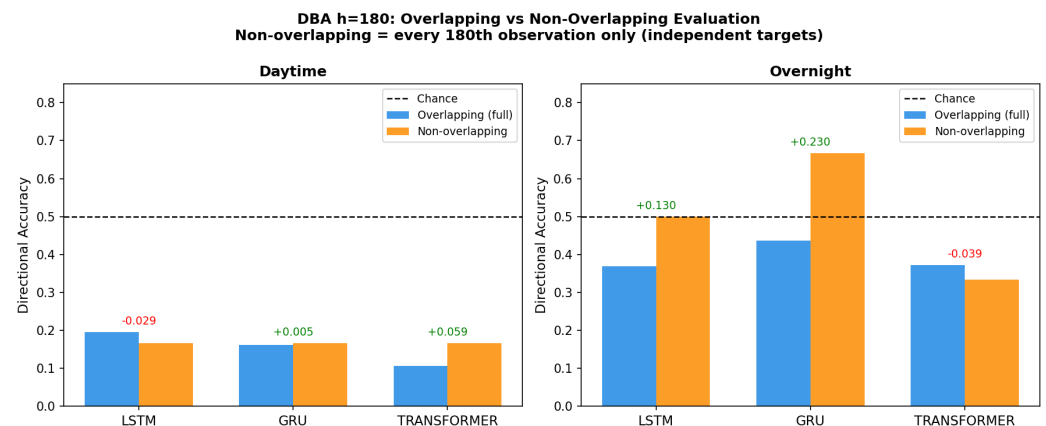


Figure 27. Comparison of directional accuracy under overlapping and non-overlapping target construction for DBA daytime $h = 180$. Non-overlapping DA (every h -th observation) is 16.7%, compared with 10.7–19.6% for the full overlapping set. The failure persists in the non-overlapping subsample, ruling out the overlapping target structure as the primary cause of the catastrophic directional misalignment.

6. Discussion

The results reveal a consistent and economically significant pattern: predictability in commodity ETF returns is highly selective, emerging in specific (asset, session, horizon) combinations rather than being a broad feature of this asset class. Three findings stand out. First, DBP overnight returns are robustly predictable by all three deep learning architectures at medium and long horizons, with directional accuracy reaching 77% and annualized Sharpe ratios reaching 15 at the 180-day horizon. Second, DBB daytime at 30 days and DBO daytime at 180 days offer more modest but statistically supported signals, particularly for the Transformer. Third, the majority of (ETF, session, horizon) configurations are near-random, providing evidence consistent with weak-form efficiency for most commodity price series when the return is properly conditioned on session and forecast horizon.

6.1. Precious Metals Overnight Predictability: Microstructure Rationale

The pronounced overnight predictability of DBP is interpretable within a microstructure framework. The overnight price gap in a commodity ETF, the difference between today's opening and yesterday's closing price reflect information that accumulated while the domestic equity market was closed: after-hours news flow, offshore futures trading (notably the London Metal Exchange for gold and silver), and macro information that arrives outside of regular trading hours. The feature set constructed here, comprising lagged daytime and overnight returns at multiple horizons for all six commodities ETFs, capture cross-market information flow that is systematically embedded in the opening price adjustment of a precious metals ETF, but that is not immediately impounded into the closing price. This interpretation is consistent with the theoretical framework of [45] on gradual information diffusion across related asset markets. [46] provides direct microstructural evidence for this channel: using intraday data for gold futures, they show that the largest permanent price adjustments occur in the overnight session when the London fix is absent, precisely the setting where informational asymmetries are greatest and where multi-period momentum features of the kind used here are most likely to retain predictive power. Their finding that overnight returns lead daytime returns over multi-week horizons supports the low-frequency overnight signal documented in Tables 2 and 3.

The session asymmetry itself, strong overnight predictability versus near-zero daytime predictability for the same ETF, reinforces this interpretation. Daytime trading in a precious metals ETF involves continuous price discovery under high volume and transparency, conditions under which momentum signals are rapidly arbitrated away by informed participants. The overnight gap, by contrast, aggregates a longer interval of illiquid, opaque price adjustment, creating a window in which systematic patterns can persist long enough to be predictable from recent return history. The fact that all three architectures identify this pattern with similar accuracy at $h = 180$ (LSTM 76.7%, GRU 76.5%, Transformer 76.7%) suggests the signal is robust to modeling choice, not an artifact of any single architecture.

6.2. Base Metals and Oil: Horizon-Specific Predictability

The predictability of DBB daytime returns at the 30-day horizon, concentrated in the Transformer, and of DBO daytime at the 180-day horizon is consistent with the distinct economic drivers of base metals and oil at different forecast horizons. Base metals price dynamics at the 30-day horizon are heavily influenced by inventory cycles, Chinese manufacturing activity, and short-run supply disruptions, which are likely to generate persistent sequential patterns that the Transformer's attention mechanism is better equipped to model than the GRU's fixed hidden-state capacity. At longer horizons, these shorter cycles are washed out, and the accuracy for DBB fades substantially (GRU falls to chance). Oil at the 180-day horizon, by contrast, is primarily driven by supply capacity, OPEC policy, and macroeconomic conditions, low-frequency drivers that generate persistent multi-month trends, which LSTM and the Transformer capture effectively.

The failure of the GRU relative to LSTM and Transformer on the key predictive configurations warrants attention. The GRU lags the Transformer by a substantial margin on DBB daytime at $h = 30$ (50.9% vs. 64.3%) and falls below the LSTM on DBO daytime at $h = 180$ (Sharpe 6.78 vs. 8.27). A plausible explanation is that the GRU's two-gate design creates an information bottleneck at longer sequences: the combined reset-update mechanism must simultaneously decide how much past state to carry forward and how much new input to incorporate, whereas the LSTM's separate input, forget, and output gates provide more flexible routing of long-range dependencies, and the Transformer's self-attention directly accesses any position in the input sequence. When the relevant signal spans 30 or more trading days of cross-asset return history, the GRU's bottleneck produces an information loss that degrades directional accuracy in exactly the configurations where the other two architectures succeed.

6.3. Implications for Market Efficiency and Practical Use

The findings admit two interpretations that are not mutually exclusive. Under the *risk-premium* interpretation, the overnight Sharpe ratio of 15 for DBP at $h = 180$ reflects compensation for bearing illiquidity and informational risk in the overnight session of a derivative product (the ETF) that itself carries a small tracking-error discount relative to the underlying metals. Under the *inefficiency* interpretation, at least some of the signal represents actionable mispricing that survives to the next open, which is consistent with the observed result that simple closing-opening long-short strategies applied to the ETF would have been profitable across the test period. Distinguishing between these interpretations empirically requires detailed transaction-cost modeling (bid-ask spreads, settlement lags for ETF overnight strategies, and the impact of position sizing on the instrument liquidity), which is beyond the scope of this paper. The Sharpe ratios reported here are pre-cost and should be interpreted as upper bounds on implementable performance.

The scope of this paper's efficiency claim is deliberately conditional: the findings suggest that deep learning models applied to a specific feature set (six commodity ETF return

series decomposed by session) detect predictability in a small subset of (asset, session, horizon) configurations over the sample period. They do not imply that commodity markets in general are inefficient or that the identified patterns will persist out-of-sample, particularly in light of the documented regime sensitivity of the DBA results (Section 5.11) and the limited degree to which the identified predictability at long horizons ($h = 180$) has been separately validated against truly out-of-sample data (post-2025). We are further unable to fully rule out overfitting as a partial contributor to the magnitude of the reported Sharpe ratios at long horizons, despite the use of walk-forward validation and seed sensitivity analysis, given the structural difficulty of achieving power in a dataset that spans fewer than 10 non-overlapping 180-day windows on the true test set.

6.4. Directions for Future Research

Several extensions follow directly from the results. First, extending the feature set to include macroeconomic and sentiment variables, term structure of interest rates, dollar index, global purchasing managers' indices, commodity-specific news sentiment scores, would test whether the deep learning signal is subsumed by, or complementary to standard macro predictors. Second, a richer regime-conditioning layer, whether implemented through a hidden Markov model, a regime-switching network, or an attention-weighted contextual embedding could address the DBA regime failure documented in Section 5.11 by explicitly conditioning directional predictions on detected market states. Third, the return decomposition framework applied here to ETFs could be applied directly to commodity futures, where the overnight session corresponds to after-hours electronic trading, and where microstructure data (bid-ask spread, order imbalance) is available to test the informational interpretation of the overnight signal more directly. Fourth, the architecture comparison presented here is limited to three sequence models, including attention mechanisms with relative positional encoding and multi-resolution convolutional feature extractors would allow more fine-grained analysis of the structural properties of the sequence models that drive predictability at different horizons.

7. Conclusions

Return decomposition into daytime ($\log C_t/O_t$) and overnight ($\log O_t/C_{t-1}$) components is the central methodological contribution of this paper. By separating within-session price discovery from overnight information adjustment, the framework identifies a clean, reproducible unit of analysis that aligns with distinct microstructure regimes, high-volume continuous price discovery during trading hours versus illiquid, information-rich gap adjustment before the open. The results demonstrate that this decomposition is not merely a technical detail but is empirically consequential: the most striking results in the paper (DBP overnight predictability) arise precisely in the session where informational asymmetries between domestic and offshore markets are largest, while daytime returns for the same ETF are near-random.

This paper has examined whether LSTM, GRU, and Transformer architectures can produce economically meaningful directional predictions for the returns of six commodity ETFs (DBA, DBB, DBC, DBE, DBO, DBP) over forecast horizons of 1 to 180 trading days, using a 20-day sequence of cross-ETF daytime and overnight returns as input features. The analysis spans January 2007 to December 2025, with a strict chronological 80/20 train-test split and expanding walk-forward validation across three independent test folds.

What we learn: Predictability in commodity ETF returns is highly selective and horizon-specific. The key findings are:

- **DBP overnight returns** are robustly predictable at medium and long horizons. All three architectures achieve directional accuracy of 64–67% at $h = 30$ and 77% at

$h = 180$ (Sharpe 15.1), with results statistically significant after HAC correction and FDR adjustment.

- **DBB daytime at $h = 30$ and DBO daytime at $h = 180$** offer more moderate but supported signals, concentrated in the Transformer and LSTM, respectively. The Transformer's self-attention mechanism provides meaningful advantages over recurrent models in these configurations.
- **Agricultural ETF returns (DBA) at $h = 180$** display systematic directional inversion rather than unpredictability, a regime-specific failure in which momentum features trained before 2021 produce contrarian predictions systematically during the post-COVID commodity price cycle.
- **Robustness:** Results for DBP overnight at $h = 60$ and $h = 180$ survive three-fold walk-forward validation (fold-average DA 73.2% and Sharpe 9.84), FDR correction across 180 tests, Diebold–Mariano tests with HLN correction, and seed and hidden-size sensitivity sweeps. The DBO daytime $h = 180$ configuration also passes walk-forward validation with fold-average DA 68.2%.
- **Linear baselines** (logistic regression, ridge regression, momentum) are outperformed by deep learning on the headline configurations, particularly DBB daytime at $h = 30$ (Transformer DA 69.9% vs. logistic 60.9%) and DBP overnight at $h = 180$ (Transformer Sharpe 15.1 vs. ridge 9.89), confirming that non-linear sequence modelling adds incremental value beyond the best linear alternative.

This paper has several limitations that should be considered when interpreting the results. The headline Sharpe ratios are pre-transaction-cost and should be taken as upper bounds; overnight ETF strategies involve early-morning market orders whose bid-ask spreads and market impact could materially reduce net returns, particularly at long horizons where position turnover is infrequent but position sizes are larger. The HAC standard errors, inflated by a factor of approximately 7.4 at $h = 180$ relative to i.i.d. standard errors, address serial correlation in the overlapping return series, but the resulting tests have low power in a test set that spans fewer than 10 non-overlapping 180-day windows; formal Diebold–Mariano tests with HLN correction confirm significance for the primary configurations, but the inference remains fragile at the extreme horizon. The feature set, while comprehensive across six ETFs and two sessions, excludes macroeconomic series, commodity-specific supply/demand data, and sentiment signals that could either sharpen the identified signals or absorb them. The walk-forward results for DBP overnight are somewhat sensitive across folds (DA standard deviation 0.21–0.38 at $h = 60$ and $h = 180$), implying that while the mean directional accuracy is high, individual fold estimates vary substantially. Finally, the DBA regime diagnosis, while providing a plausible explanation for the catastrophic failure at $h = 180$, is ex post and cannot be used to argue that a real-time regime-conditioning strategy would have performed well without additional validation on held-out data.

Notwithstanding these limitations, the headline Sharpe ratio of 15.1 for DBP overnight at $h = 180$ warrants contextualization before it is taken at face value. The figure is an out-of-sample realized Sharpe computed on the held-out test set, not a training-set artefact. Three-fold expanding walk-forward validation confirms persistence across time: the Transformer records a fold-average Sharpe of 10.86 at $h = 180$, with the across-fold standard deviation of 22.4 reflecting the very small number of non-overlapping 180-day windows within each three-year fold rather than genuine regime instability. The moving-block bootstrap 95% confidence interval of 7.5–45.3 further anchors the result: even at the conservative lower bound, the risk-adjusted payoff remains economically significant. Mechanistically, the DBP overnight signal is anchored to the well-documented concentration of permanent precious-metals price adjustment during non-trading hours when the London

fix is absent [46], providing an economic rationale for persistence that distinguishes this finding from purely data-driven factor strategies which lack a corresponding microstructure mechanism. Practitioners should nonetheless expect materially lower forward-looking Sharpe ratios after accounting for post-publication alpha decay, capacity constraints, and the possibility of structural regime change in commodity markets; the results here are best interpreted as establishing the existence and direction of a predictable overnight pattern, not as a direct trading-performance forecast. We hope this paper contributes to the growing literature on the joint role of session-specific microstructure and non-linear sequence modeling in commodity price predictability, and that the return decomposition framework introduced here proves useful for future studies in related asset classes.

Author Contributions: Conceptualization, S.P., T.K. and E.P.; methodology, S.P. and T.K.; software, S.P.; validation, S.P. and T.K.; formal analysis, S.P.; investigation, S.P. and T.K.; data curation, S.P.; writing—original draft preparation, S.P.; writing—review and editing, T.K. and E.P.; visualization, S.P.; supervision, E.P.; project administration, E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available from Yahoo Finance (<https://finance.yahoo.com>) and are retrievable using standard Python libraries (e.g., `yfinance`). The trained model weights, processed feature matrices, and evaluation outputs are available from the corresponding author upon reasonable request. The analysis code is available at <https://github.com/triparnanalytics/Commodities> (last accessed 19 July 2026).

Acknowledgments: The authors thank seminar participants for helpful comments on earlier drafts of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gorton, G.; Rouwenhorst, K.G. Facts and Fantasies about Commodity Futures. *Financ. Anal. J.* **2006**, *62*, 47–68. [CrossRef]
2. Tang, K.; Xiong, W. Index Investment and the Financialization of Commodities. *Financ. Anal. J.* **2012**, *68*, 54–74. [CrossRef]
3. Fama, E.F. Efficient Capital Markets: A Review of Theory and Empirical Work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
4. Lo, A.W.; MacKinlay, A.C. Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. *Rev. Financ. Stud.* **1988**, *1*, 41–66. [CrossRef]
5. Brock, W.; Lakonishok, J.; LeBaron, B. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *J. Financ.* **1992**, *47*, 1731–1764. [CrossRef]
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
7. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
8. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
10. Fischer, T.; Krauss, C. Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]
11. Gu, S.; Kelly, B.; Xiu, D. Empirical Asset Pricing via Machine Learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
12. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
13. Jiang, W. Applications of Deep Learning in Stock Market Prediction: Recent Progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
14. Nabipour, M.; Nayyeri, P.; Jabani, H.; Mosavi, A.; Salwana, E.; S., S. Deep Learning for Stock Market Prediction. *Entropy* **2020**, *22*, 840. [CrossRef] [PubMed]

15. Selvin, S.; Vinayakumar, R.; Gopalakrishnan, E.A.; Menon, V.K.; Soman, K.P. Stock Price Prediction Using LSTM, RNN and CNN-Sliding Window Model. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 1643–1647. [\[CrossRef\]](#)
16. Cao, R. Stock Price Prediction Using Deep-Learning Models: CNN, RNN, and LSTM. *SHS Web Conf.* **2024**, *196*, 02004. [\[CrossRef\]](#)
17. Mehtab, S.; Sen, J. Stock Price Prediction Using CNN and LSTM-Based Deep Learning Models. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 447–453. [\[CrossRef\]](#)
18. Lou, D.; Polk, C.; Skouras, S. A Tug of War: Overnight Versus Intraday Expected Returns. *J. Financ. Econ.* **2019**, *134*, 192–213. [\[CrossRef\]](#)
19. Erb, C.B.; Harvey, C.R. The Strategic and Tactical Value of Commodity Futures. *Financ. Anal. J.* **2006**, *62*, 69–97. [\[CrossRef\]](#)
20. Jegadeesh, N.; Titman, S. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *J. Financ.* **1993**, *48*, 65–91. [\[CrossRef\]](#)
21. Moskowitz, T.J.; Ooi, Y.H.; Pedersen, L.H. Time Series Momentum. *J. Financ. Econ.* **2012**, *104*, 228–250. [\[CrossRef\]](#)
22. Asness, C.S.; Moskowitz, T.J.; Pedersen, L.H. Value and Momentum Everywhere. *J. Financ.* **2013**, *68*, 929–985. [\[CrossRef\]](#)
23. Vukovic, D.B.; Ingenito, S.; Maiti, M. Time Series Momentum: Evidence from the European Equity Market. *Heliyon* **2023**, *9*, e12989. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Waggle, D.; Agrawal, P. Is the “Sell in May and Go Away” Adage the Result of an Election-Year Effect? *Manag. Financ.* **2018**, *44*, 1070–1082. [\[CrossRef\]](#)
25. Kaastra, I.; Boyd, M. Designing a Neural Network for Forecasting Financial and Economic Time Series. *Neurocomputing* **1996**, *10*, 215–236. [\[CrossRef\]](#)
26. Guney, O.B.; Saichandran, K.S.; Elzokm, K.; Zhang, Z.; Kolachalama, V.B. Active Feature Acquisition via Explainability-Driven Ranking. *Proc. Mach. Learn. Res.* **2025**, *267*, 20748–20765.
27. Rompotis, G.G. Interfamily Competition on Index Tracking: The Case of the Vanguard ETFs and Index Funds. *J. Asset Manag.* **2009**, *10*, 263–278. [\[CrossRef\]](#)
28. Farinella, J.A.; Kubicki, R. The Performance of Exchange Traded Funds and Mutual Funds. *J. Account. Financ.* **2018**, *18*. [\[CrossRef\]](#)
29. Kundu, T.; Pinsky, E. Predicting Daily Stock Price Directions with Deep Learning Models. *Mach. Learn. Appl.* **2025**, *22*, 100744. [\[CrossRef\]](#)
30. Zhang, X.; Pinsky, E. S&P-500 vs. Nasdaq-100 Price Movement Prediction with LSTM for Different Daily Periods. *Mach. Learn. Appl.* **2025**, *19*, 100617. [\[CrossRef\]](#)
31. Beniwal, M.; Singh, A.; Kumar, N. Alternative to Buy-and-Hold: Predicting Indices Direction and Improving Returns Using a Novel Hybrid LSTM Model. *Int. J. Artif. Intell. Tools* **2023**, *32*, 2350028. [\[CrossRef\]](#)
32. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015. [\[CrossRef\]](#)
33. Aboody, D.; Even-Tov, O.; Lehavy, R.; Trueman, B. Overnight Returns and Firm-Specific Investor Sentiment. *J. Financ. Quant. Anal.* **2018**, *53*, 485–505. [\[CrossRef\]](#)
34. Jurdi, D.J. Intraday Jumps, Liquidity, and U.S. Macroeconomic News: Evidence from Exchange Traded Funds. *J. Risk Financ. Manag.* **2020**, *13*, 118. [\[CrossRef\]](#)
35. Jansen, W.J. Are Commodity Markets Conditionally Efficient? *Manag. Financ.* **2020**, *46*, 981–996.
36. Agrawal, P.; Clark, J.M.; Agarwal, R.; Kale, J.K. An Inter-temporal Study of ETF Liquidity and Underlying Factor Transition, 2009–2014. *J. Trading* **2014**, *9*, 69–78. [\[CrossRef\]](#)
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015. [\[CrossRef\]](#)
38. Quantified Strategies. *Commission Trading: What Is Realistic to Pay?* Quantified Strategies: New York, NY, USA, 2025.
39. Newey, W.K.; West, K.D. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **1987**, *55*, 703–708. [\[CrossRef\]](#)
40. Hansen, L.P.; Hodrick, R.J. Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis. *J. Political Econ.* **1980**, *88*, 829–853. [\[CrossRef\]](#)
41. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [\[CrossRef\]](#)
42. Harvey, D.I.; Leybourne, S.J.; Newbold, P. Testing the Equality of Prediction Mean Squared Errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [\[CrossRef\]](#)
43. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* **1995**, *57*, 289–300. [\[CrossRef\]](#)
44. Pattnaik, S.; Jain, C.; Pinsky, E. Spatiotemporal Return Decomposition and Multi-Strategy Performance Analysis in Dow Jones Industrial Average Constituents: A 20-Year Empirical Investigation. *Int. J. Financ. Stud.* **2026**, *14*, 145. [\[CrossRef\]](#)

45. Hong, H.; Torous, W.; Valkanov, R. Do Industries Lead Stock Markets? *J. Financ. Econ.* **2007**, *83*, 367–396. [[CrossRef](#)]
46. Jurdi, D.J. Predicting the US Market Return: Evidence from the Gold Market. *Int. Rev. Financ. Anal.* **2023**, *86*, 102531. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.