Summary of Model Performance Comparison

This document provides a comparative analysis of four distinct modeling approaches—QR, GARCH, DNN and LSTM—in the context of Value-at-Risk (VaR) estimation. The evaluation focuses on their performance across different confidence levels (α = 1%, 5%, 10%) based on several statistical tests and loss metrics.

Note on Experimental Setup:
Based on the available daily data, a feature selection process was conducted by evaluating multiple feature combinations. The feature sets used for each model represent the best-performing configurations among those empirically tested.

Model Performance:
1. QR (LASSO) demonstrates robust performance at higher confidence levels (5% and 10%), passing both the Dynamic Quantile (DQ) and Kupiec tests, while maintaining a notably low average tick loss across all levels. However, at the 1% level, it fails the DQ test despite an acceptable failure rate.
2. DNN (MLP) exhibits results similar to QR, passing both tests at the 5% and 10% levels but failing the DQ test at the 1% level. Its tick loss, however, is slightly higher than that of QR.
3. LSTM yields mixed outcomes: it passes both tests at the 5% level but fails the DQ test at the 1% and 10% levels. While its failure rates remain reasonable, its average tick loss is somewhat higher compared to QR and DNN.
4. GARCH shows exceptional performance at the 1% level, passing all tests. However, at the 5% and 10% levels, it fails the Kupiec test due to overly conservative VaR estimates (failure rates too low), resulting in substantially higher average tick losses compared to other models.

Overall Implications:
1. The results highlight a trade-off between model complexity and regulatory backtest performance. While the traditional GARCH excels at extreme quantiles (α=1%), it tends to overestimate risk at moderate confidence levels. ML/DL models (QR, DNN, LSTM) offer competitive forecast accuracy, particularly in terms of average quantile loss, yet they may struggle to satisfy the rigorous conditional

coverage properties required by the DQ test in certain tail scenarios. Therefore, model selection should be guided by the specific risk management objective—whether the priority is to excel in statistical backtests at extreme confidence levels or to minimize forecast error over more common risk thresholds.

2. For tasks with limited sample sizes and relatively simple data structures, simpler approaches such as LASSO regression can perform on par with more complex DNN and LSTM models.

Potential Avenues for Enhancing VaR Forecasts:

1. Explore hybrid methodologies that combine GARCH with machine learning or deep learning models, with the aim of mitigating the clustering of violations observed in DNN and LSTM at the $\alpha=1\%$ level.
2. Expand the sample size, for instance by utilizing high-frequency data to forecast intraday VaR, or by extending the temporal scope of the dataset.
3. Incorporate additional predictive features, such as the VIX index, trading volume, bid-ask spreads, or sentiment measures derived from textual data (e.g., Federal Reserve meeting minutes).
4. Employ GNN models to analyze volatility spillover effects from related assets (e.g., crude oil, gold, cryptocurrencies) on the VaR of the target asset.