Generative AI and LLM Ensembles for Forecasting Capital Markets

Deep Learning and AI Conference
Silicon Valley

Yigal Jhirad
Emanuel Scoullos
Siddharth Samsi

NVIDIA GTC
March 19, 2024

GTC 2024: Table of Contents

Generative AI and LLM Ensembles

Machine Learning Landscape

Challenges in Financial Time Series Forecasting: Small Data

Traditional Deep learning vs. Deep Learning for Finance

Model Implementation

- Transformers Architecture
- Training data
- Training Workflow
- Forecasting Results

Summary

Biographies

Yigal Jhirad

Emanuel Scoullos, Ph.D

Siddharth Samsi, Ph.D

Special Thanks to Justin Hodgsen

DISCLAIMER: This presentation is for information purposes only. The presenter accepts no liability for the content of this presentation, or for the consequences of any actions taken on the basis of the information provided. Although the information in this presentation is considered to be accurate, this is not a representation that it is complete or should be relied upon as a sole resource, as the information contained herein is subject to change.

GTC 2024: Artificial Intelligence

Machine Learning

Data: Structured/Unstructured

Asset Prices, Volatility
Fundamentals (P/E, PCE, Debt to Equity)
Macro (GDP Growth, Interest Rates, Oil prices)
Technical(Momentum)
Sentiment Analysis
Security Attributes (Country, Sector, Industry)

Supervised Learning

Neural Networks
Support Vector Machines
Classification & Regression
Trees
K-Nearest Neighbors
Regression

Unsupervised Learning

LLMs/Transformers

Generative Adversarial Network
Manifold Learning
Transfer Learning
Cluster Analysis
Principal Components
Expectation Maximization

Reinforcement Learning

DQN
Q-Learning
Q-Matrix
Trial & Error

Figure 1: Yigal Jhirad

- **Inflation**
  - Inflation erodes purchasing power
  - While a moderate amount of inflation may be a sign of a healthy economy, too low or too high can have broader economic consequence
  - Modeling inflation is difficult with a limited number of regimes and uneven impact across time with varying time horizons

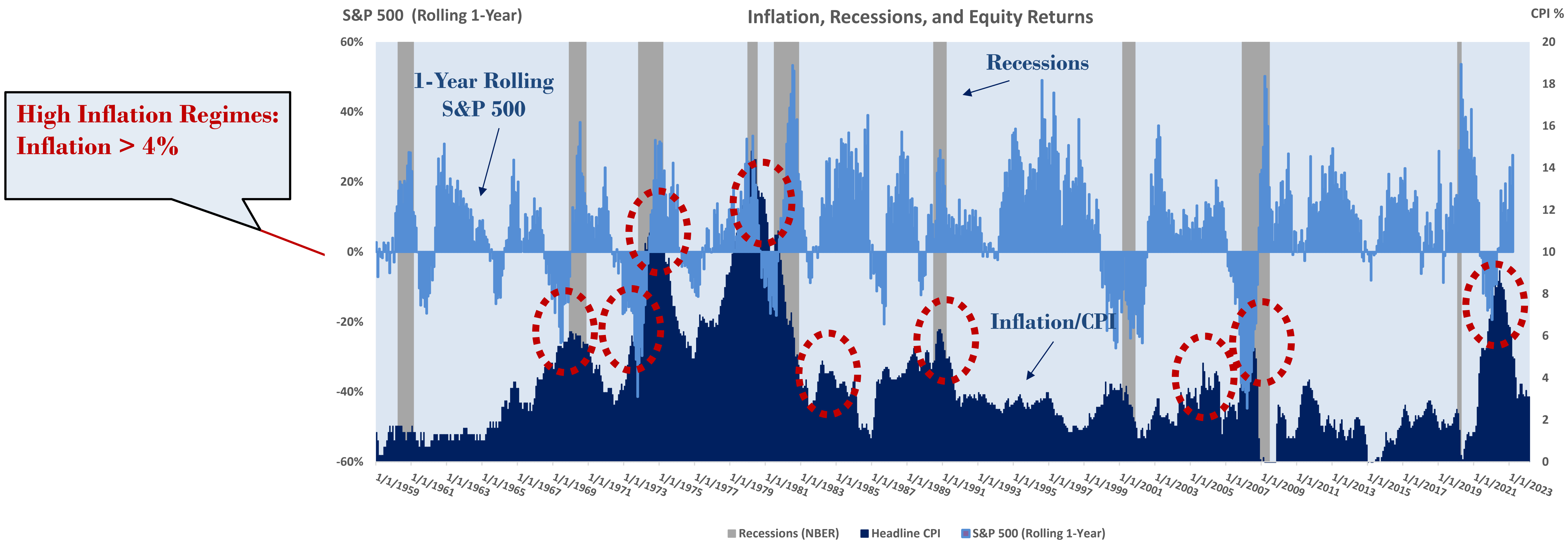
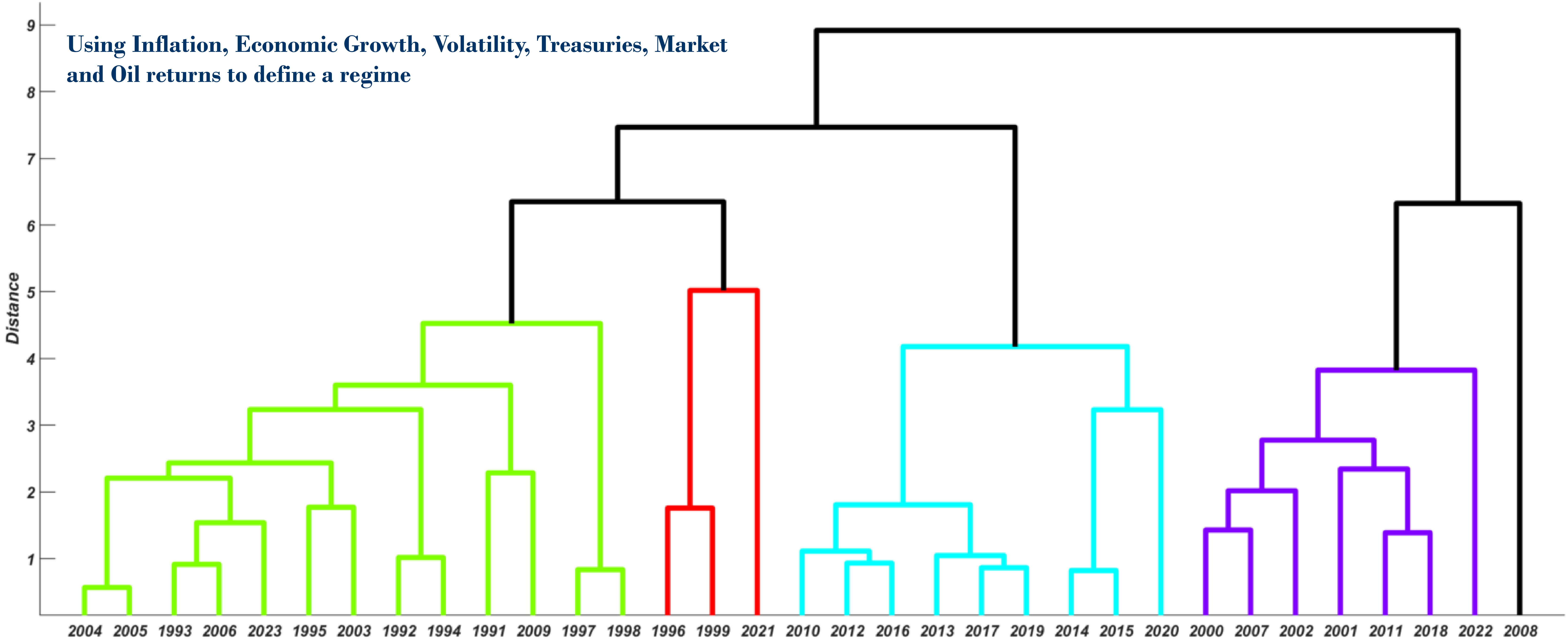


Figure 2: Bloomberg, NBER

GTC 2024: Small Data - Regimes

Regimes – Cluster Analysis



GTC 2024: Deep Learning in Finance

Traditional Deep Learning

- Big data: Millions to Billions of examples
- Focused on image and text data
- Stationary data
- LLMs, MoE

Deep Learning for Finance

- Small data: Hundreds to thousands of examples
  - Easier to Overfit
  - Noise/Pattern recognition
- Focused on tabular data
- Stationary and Non-stationary data
- Synchronization - Input lag between policy decisions and market impact

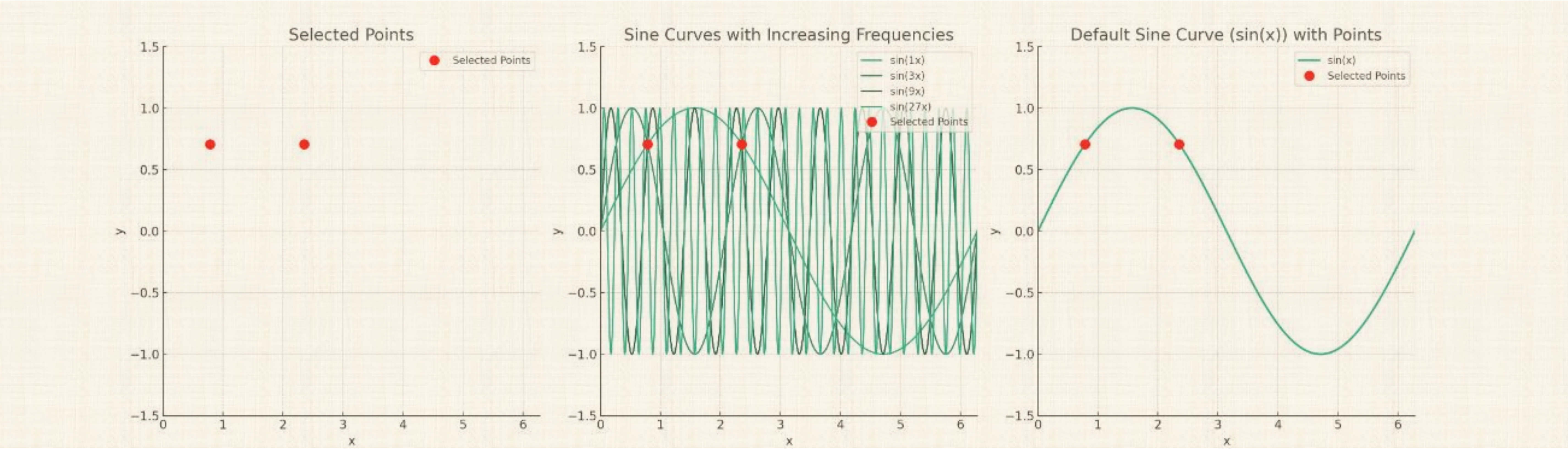
GTC 2024: Model Strategy

- **Preprocess Inputs**
  - — Limit number of inputs to most meaningful and relevant data
  - — Based on heuristics or classical quantitative techniques
- **Constrain Solution Set**
  - — Reduce solution space so be more congruent with input parameters and scope of data
  - — Dimensionality reduction techniques (e.g. Regularization, Manifold Learning)
  - — Avoid spurious outcomes
- **Ensemble Models**
  - —Leverage deep learning approaches including LLMs and classical quantitative models
  - —Diversify signals and capture nonlinear relationships. Comparative advantage across models including LLMs
  - —Chaos Theory– avoid oversensitivity to initial conditions and create more stability

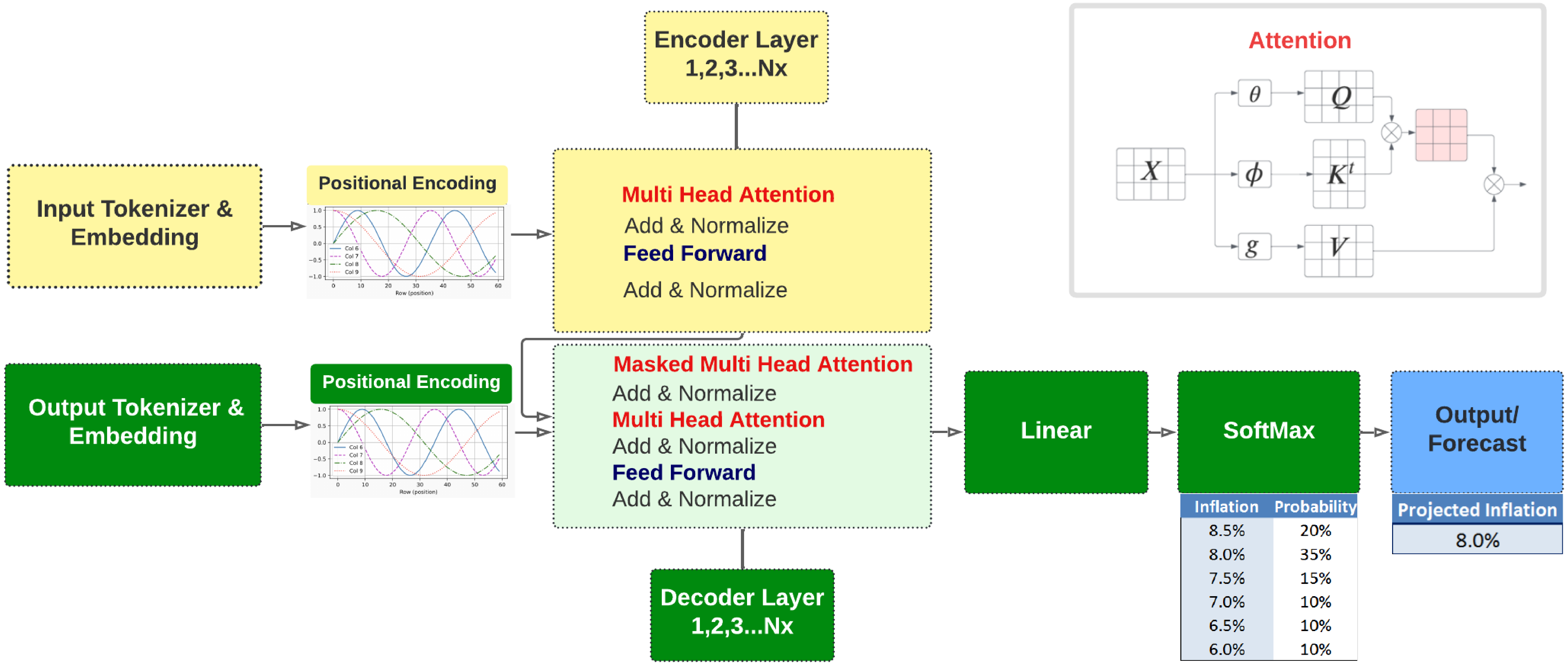
Concept	Field	Equation
Sensitivity to Initial Conditions	Chaos Theory	$d(t) \approx e^{\lambda t} d(0)$
Ensemble Averaging	Machine Learning Ensembles	$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

GTC 2024: Model Strategy

- **Pattern Recognition**
  - Do two points or prices constitute a pattern?
- **Preprocess inputs - limit number of inputs to most meaningful and relevant data**
  - Now assume that the independent input feature set is a Sinusoidal wave
  - Potentially infinite solutions
- **Constrain Solution Set**
  - Reduce solution space so be more congruent with input parameters and scope of data
  - Assume the two points have to be within the same period



GTC 2024: LLMs/Transformer Architecture



Inflation	Probability
8.5%	20%
8.0%	35%
7.5%	15%
7.0%	10%
6.5%	10%
6.0%	10%

Projected Inflation
8.0%

Figure 6: Yigal Jhirad. Based on ‘Attention Is All You Need’ by Vaswani et al.

GTC 2024:LLMs/Transformers for Time Series Forecasting

- **LLMs have shown impressive performance in NLP and Vision domains and increasingly in time series applications**

- **LLMs are based on the Transformer architecture which features that lend themselves to certain problems in quantitative finance particularly the self-attention mechanism designed to learn context and relationships across inputs**
  - Similar to measuring relationships in financial data using covariance or correlation
- **Poincaré Recurrence – systems will return to states close to the initial states - provides an analogous framework for understanding cyclical nature of financial data.**
  - LLM's may more effectively capture these cyclical patterns and long term trends through self-attention
  - The "information volume" of the data may be retained. Certain patterns or features in the data are transformed dynamically may lead to outputs that may be recurrent under certain conditions.

Training Data

- **Monthly broad market and sector indexes, oil, inflation, volatility, term structure, money supply data from 08/2007 – 01/2024**
  - —195 Dates, Over 20 Features
- **Training data spanned 10/2007 to 08/2021**
- **Test data from 09/2021 to 01/2024**

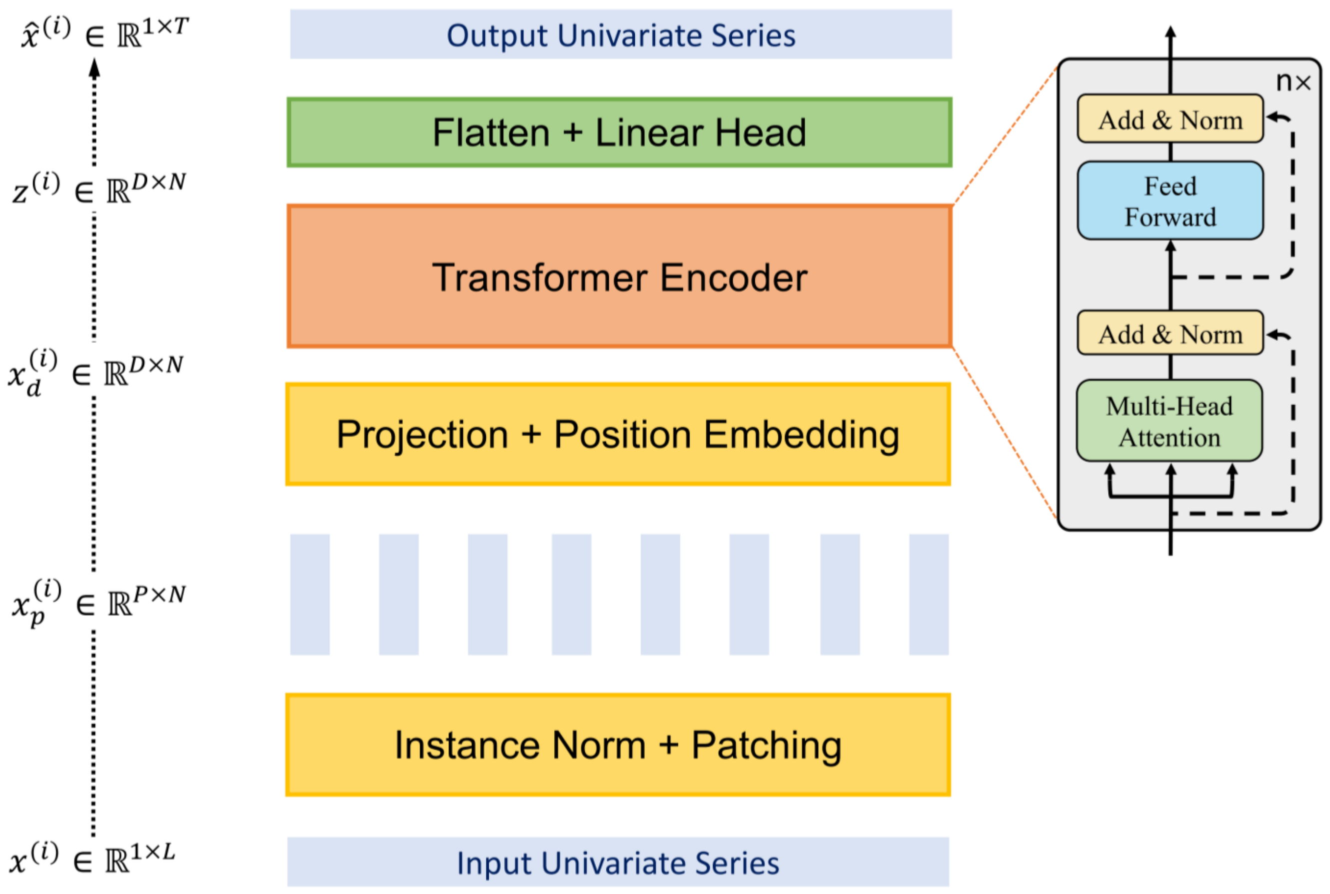
Date	Oil	Copper	Copper Price	S&P500 Materials	3M-Treasury	5YR Treasury
20081031	-0.32893	-0.36471	1.93808	-0.22177	0.436	2.8277
20081128	-0.20846	-0.11236	1.72032	-0.11219	0.0406	1.9144
20081231	-0.22798	-0.14387	1.47282	-0.00757	0.0761	1.5489
20090130	-0.21137	0.04149	1.53393	-0.07239	0.2262	1.8751
20090227	-0.04026	0.03915	1.59399	-0.08846	0.2465	1.9839
20090331	0.09108	0.20436	1.91973	0.1493	0.2009	1.6551
20090430	-0.03157	0.11304	2.13674	0.15092	0.1247	2.0104
20090529	0.2785	0.07355	2.29389	0.05549	0.1298	2.3399
20090630	0.04378	0.02753	2.35704	-0.04894	0.1775	2.5546
20090731	-0.02061	0.15726	2.7277	0.13309	0.1755	2.5144
20090831	-0.01988	0.07032	2.91952	0.0196	0.1268	2.385
20090930	0.00403	-0.00235	2.91267	0.04735	0.1075	2.3117
20091030	0.08231	0.04842	3.05371	-0.05328	0.0446	2.3084

Implementation

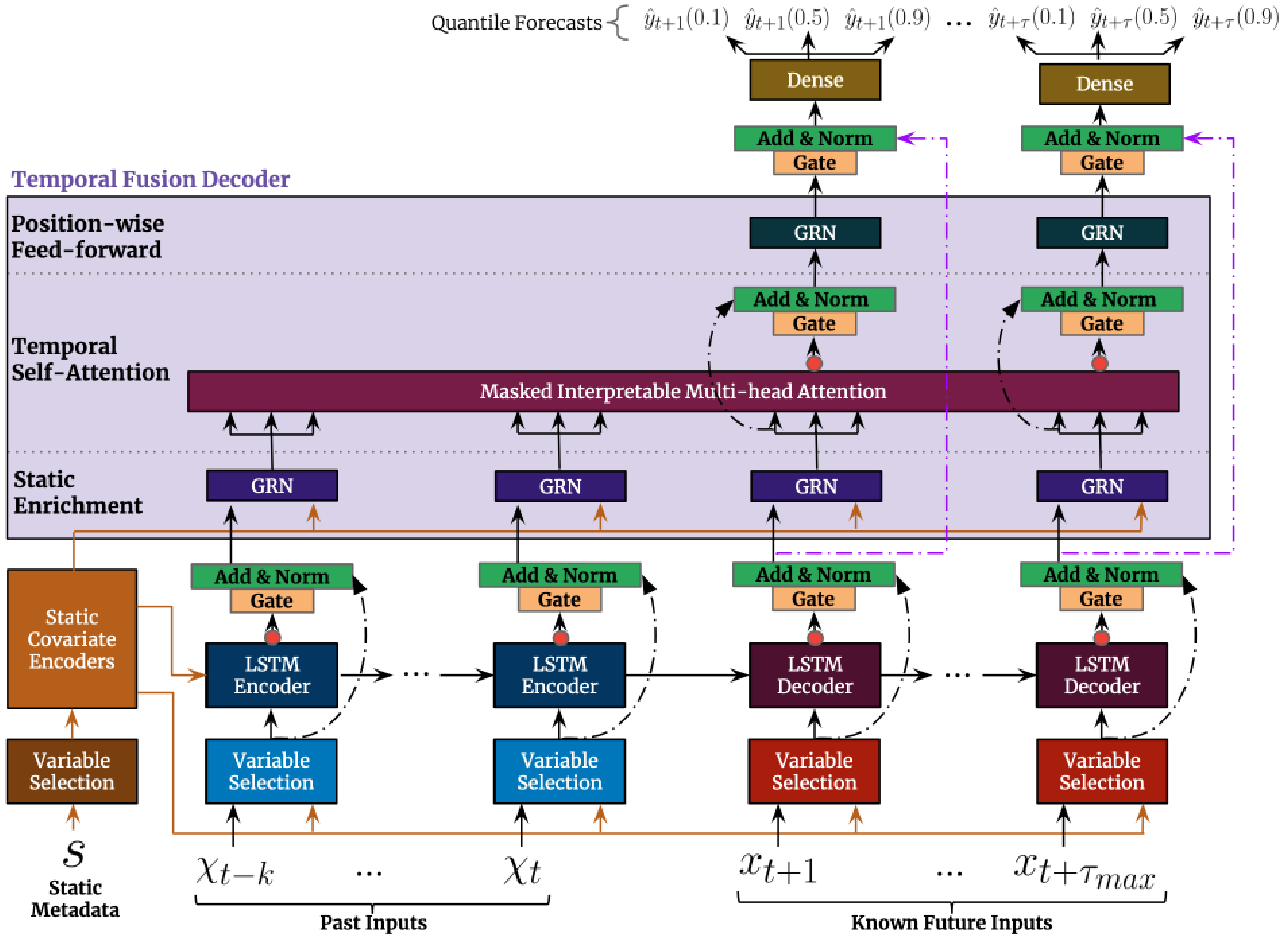
- **We use the AutoGluon[1] AutoML library for timeseries forecasting**
  - —Automated stack ensembling, deep learning for text, image, and tabular data
  - —Supports Transformer-based, deep learning and statistical models
  - —GPU accelerated[2]
- **Deep learning and Transformer models used in our example**
  - —DLinear[3]
  - —PatchTST[4]
  - —DeepAR[5]
  - —Temporal Fusion Transformer[6]
- **Statistical models used include AutoARIMA, ETS, AutoETS**

[1] AutoGluon–TimeSeries: AutoML for probabilistic time series forecasting - Shchur, Oleksandr, et al. International Conference on Automated Machine Learning. PMLR, 2023
[2] https://developer.nvidia.com/blog/advancing-the-state-of-the-art-in-automl-now-10x-faster-with-nvidia-gpus-and-rapids/
[3] Are transformers effective for time series forecasting? - Zeng, Ailing, et al., AAAI Conference on Artificial Intelligence. 2023.
[4] A Time Series is Worth 64 Words: Long-term Forecasting with Transformers - Nie, Yuqi, et al. , ICLR 2023.
[5] DeepAR: Probabilistic forecasting with autoregressive recurrent networks - Salinas, David, et al., International Journal of Forecasting. 2020.
[6] Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting - Lim, Bryan, et al., International Journal of Forecasting. 2021

Implementation: Transformer Architectures



PatchTST Architecture



Temporal Fusion Transformer Architecture

Training Workflow

- **Model training was performed on the NVIDIA DGX cloud using A100 GPUs**
- **Over 7,000 models across 84 configurations were trained**
- **Each configuration consisted of unique hyper-parameter choices and model combinations.**
  - AutoGluon first fits individual models sequentially
  - Next, the trained models are ensembled using 100 steps of the forward selection algorithm described in Caruana et al.[1]
- **Models were trained with the following losses**
  - RMSE – Root Mean Squared Error
  - MASE – Mean Absolute Scaled Error
  - SQL – Scaled Quantile Loss
  - WQL – Weighted Quantile Loss
  - Custom loss function
- **Models were trained to predict all features in the dataset**

[1] Ensemble selection from libraries of models – Caruna, et. al., Proceedings of the twenty-first international conference on Machine learning

Mean Correct Forecast Direction (Batting Average)

- **MCFD or Batting Average is closely related to market timing**
  - —Compares predicted *direction* of market movement to actual observed direction
- **Defined mathematically as:**

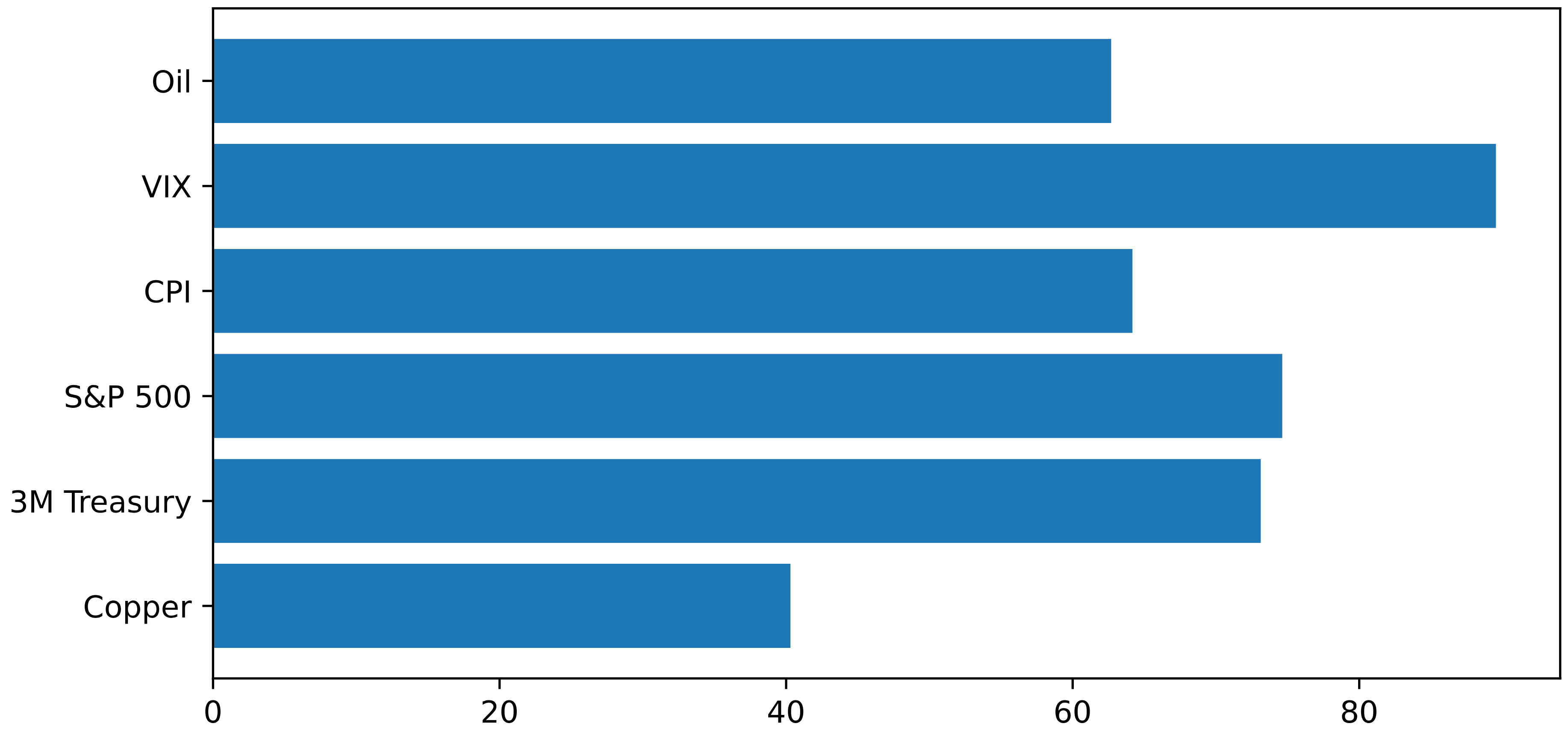
where $\vec{Y} = \vec{Y}(t+1) - \vec{Y}(t)$ **is the change in observed values**

$\vec{F} = \vec{F}(t+1) - \vec{F}(t)$ **is the change in forecast/predicted values**

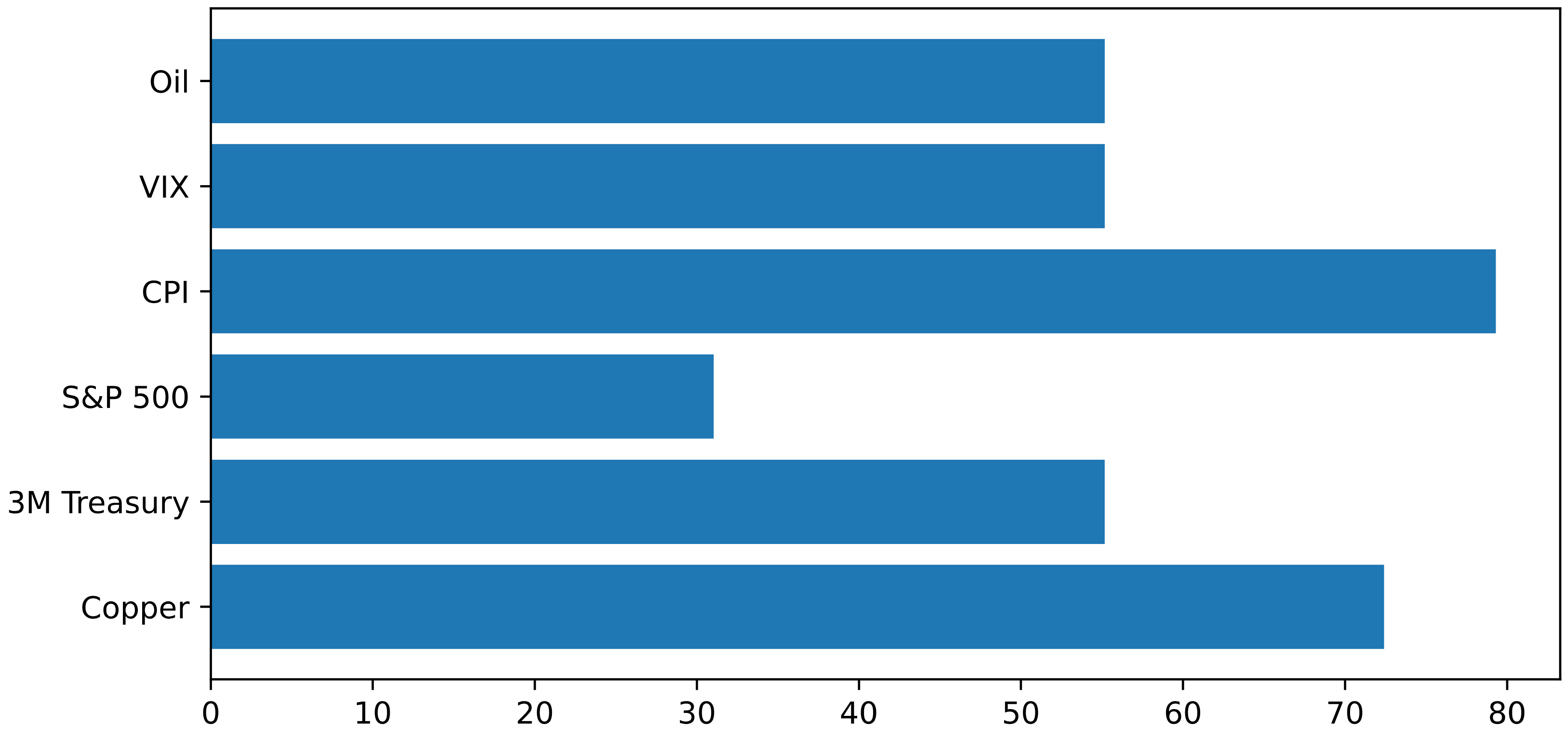
- **Batting average of predictions over the training set was used to select the best ensemble for each column in the dataset**

$$MCFD = -\frac{1}{P}\sum_{t=R}^{T}\mathbf{1}(Sign\left(\vec{Y}\right)Sign\left(\vec{F}\right) > 0)$$

Batting Average for Selected Features

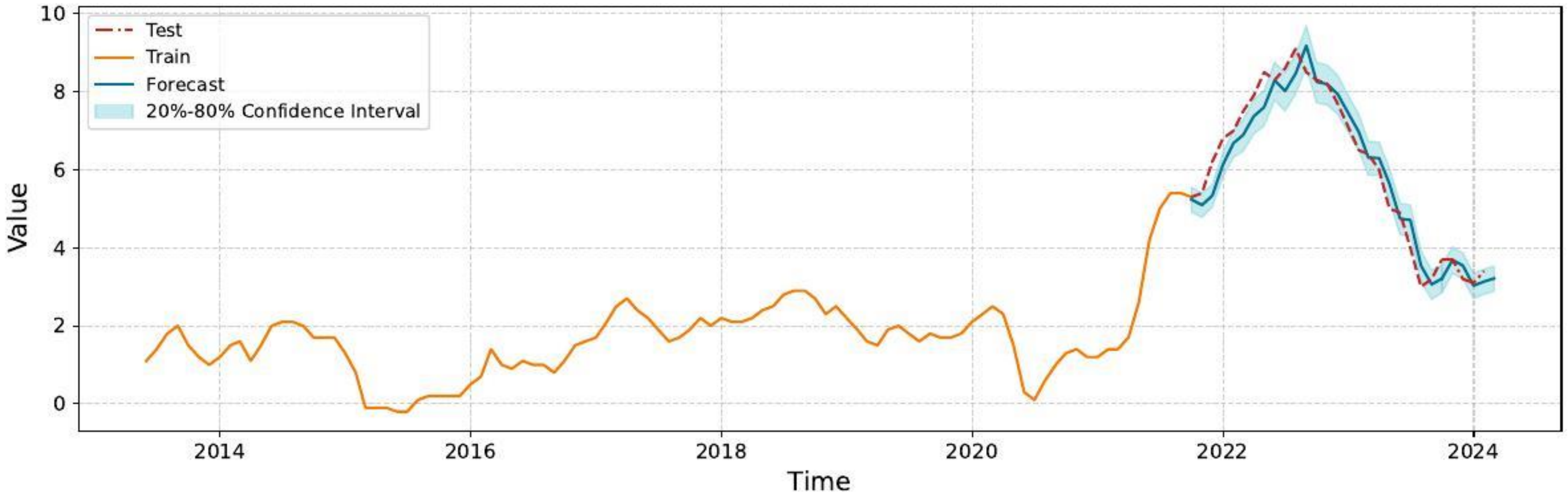


Batting Average on Train data



Batting Average on Test data

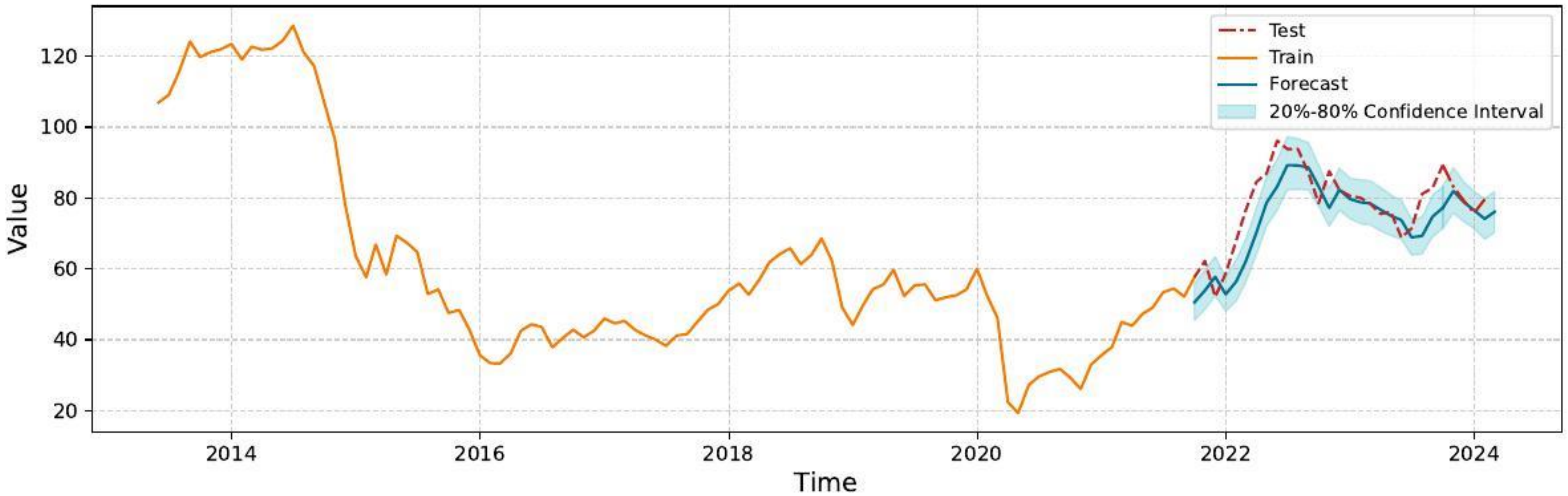
Inference results for Consumer Price Index



	Naive	AutoARIMA	DeepAR
Number of models	1	1	1
Ensemble weight	0.28	0.54	0.18

Batting Average(%)	
Train	64.2
Test	79.3

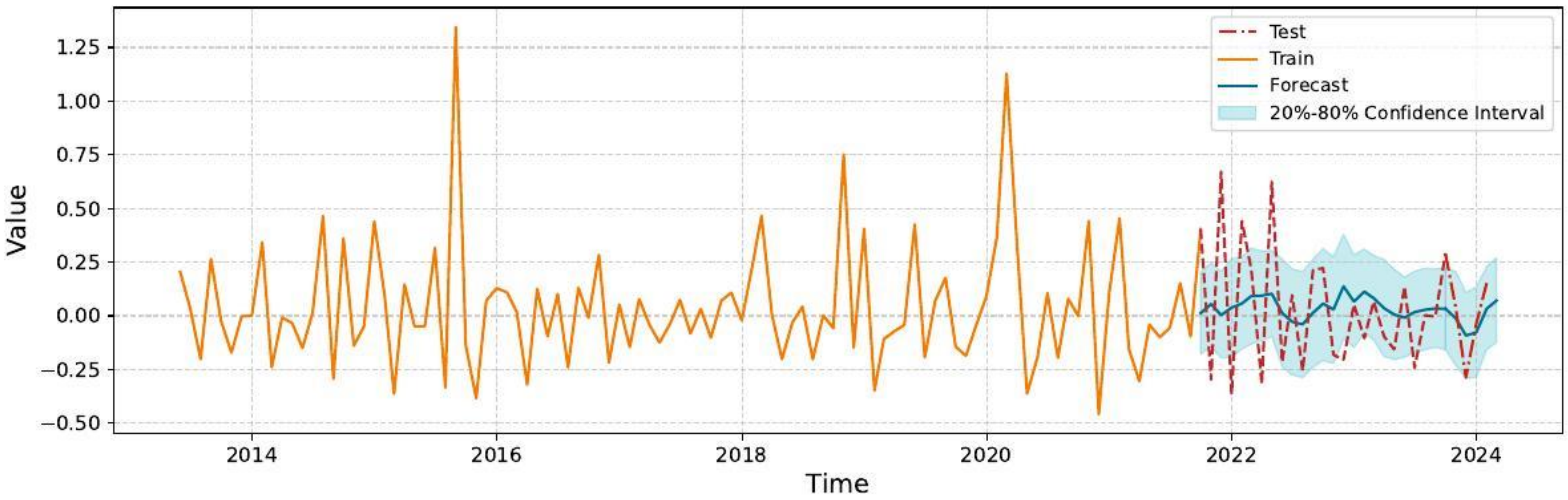
Inference results for Oil



	PatchTST	Temporal Fusion Transformer	DLinear
Number of models	4	4	3
Ensemble weight	0.30	0.40	0.30

Batting Average(%)	
Train	65.7
Test	58.6

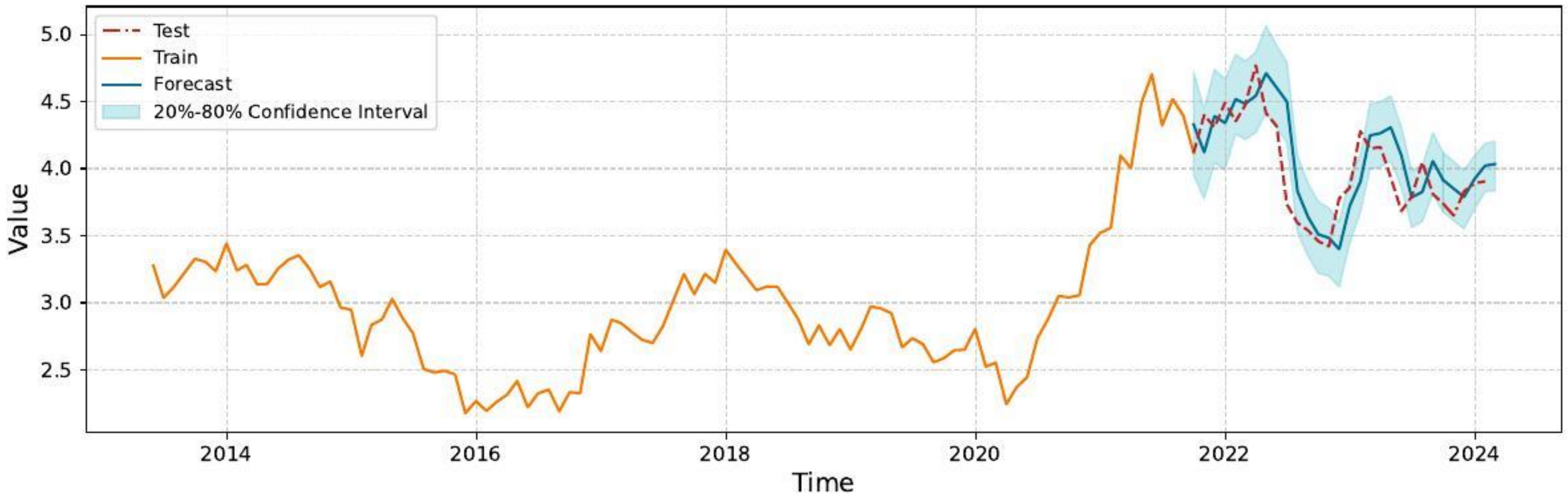
Inference results for VIX



	DeepAR	Temporal Fusion Transformer	DLinear
Number of models	17	5	1
Ensemble weight	0.78	0.21	0.01

Batting Average(%)	
Train	89.6
Test	55.2

Inference results for Copper



	Recursive Tabular	Temporal Fusion Transformer
Number of models	1	7
Ensemble weight	0.09	0.91

Batting Average(%)	
Train	40.3
Test	55.2

Inflation predictions for February

- **Inflation data was released March 12, 2024**
- **We only used end-of-January data to predict February inflation**

Date	Federal Reserve Reported CPI	Ensemble Forecast
February 2024	3.2	3.2
January 2024	3.1	3.1
December 2023	3.2	3.0
November 2023	3.7	3.5

Observations

- **Ensemble models show some auto-regressive behavior during inference**
  - This may be due to individual models in an ensemble over-fitting on the data.
- **Close examination revealed that AutoGluon's ensembling approach may need to be modified to avoid selecting models that overfit**
- **Batting Average is a useful metric for ranking ensembles but does not guarantee minimal RMSE between observations and predictions**
- **New, custom loss metrics such as the Pinball[1] loss need to be explored for improved model training**

[1] Sequential Quantile Prediction of Time Series – Gerard Biau and Benoit Patra, Information Theory, IEEE Transactions, March 2011

Summary

- **Initial exploration of transformer and deep learning ensemble models shows good potential for use with financial data.**
  - —Ensembles of models show some impressive results and promise on *small* data
- **Inference periods are limited, which can make model evaluation challenging**
  - —Need to be disciplined in preprocessing inputs and imposing constraints
- **Complements a broad quantitative modelling framework**
- **Model architecture requires significant amounts of accelerated compute**
- **Current results provide impetus for further research in this area**

Biographies

- **Yigal D. Jhirad**, Senior Vice President, is Director of Quantitative and Derivatives Strategies and Portfolio Manager for Cohen & Steers. Mr. Jhirad heads the firm's Investment Risk Committee. Prior to joining the firm in 2007, Mr. Jhirad was an executive director in the institutional equities division of Morgan Stanley, where he headed the company's quantitative and derivatives strategies effort. In previous conferences, he has presented research on Decision Trees, Neural Networks, LSTM's, Reinforcement Learning, and GAN's. Mr. Jhirad graduated Magna Cum Laude from the Wharton School of the University of Pennsylvania with a B.S. in Economics. He holds the Financial Risk Manager (FRM) designation.

  LinkedIn: https://www.linkedin.com/in/yigaljhirad/

- **Emanuel Scoullos** is a Senior Solutions Architect in the Financial Services and Technology team at NVIDIA where he focuses on GPU applications within FSI. Previously, he worked as a Data Scientist at a startup in the anti-money laundering space applying data science, analytics, and engineering techniques to construct machine learning pipelines. He earned his Ph.D. and Masters in Chemical Engineering from Princeton University and an undergraduate degree in Chemical Engineering from Rutgers University.

  LinkedIn: https://www.linkedin.com/in/emanuelscoullos/

- **Siddharth Samsi** is a Senior Solutions Architect in the Financial Services and Technology team at NVIDIA where he focuses on GPU applications within FSI. Prior to NVIDIA, he led research in distributed AI, High Performance Computing, and Energy aware computing for AI at the MIT Lincoln Laboratory Supercomputing Center. He earned his Ph.D. and Masters in Electrical and Computer Engineering from The Ohio State University.

  LinkedIn: https://www.linkedin.com/in/samsi/