Andrew Taylor EN.705.743 10/28/25

Project Proposal:" FutureFormer", using GPT to Predict the NASDAQ

Abstract

This study investigates whether a GPT-style transformer architecture can effectively model and predict high-frequency futures data in the Nasdaq-100 when trained *from scratch* in a fully supervised fashion. The core research question is: *to what extent can a GPT-style transformer capture predictive dependencies in financial time series when trained causally on bar-level futures data, and how do context length and model scale affect forecasting accuracy and information efficiency?* The dataset consists of one year of Nasdaq-100 futures formatted into fixed-duration bars and labeled using a custom triple-barrier method. The project builds custom transformer models trained on engineered numerical features under supervised learning and compares them with two baselines: ARIMA and a one-dimensional convolutional neural network. The architectures employ ALiBi positional bias, causal masking, and Flash Attention: design choices that are theoretically aligned with the temporal structure and challenge of financial forecasting. Non-stationarity is managed through rolling normalization and also periodic retraining, for which motivation will be presented instead of results given this is a snapshot study. Evaluation is conducted using accuracy, precision, and recall. A brief entropy analysis using Shannon entropy and sample entropy measures the informational richness of my chosen feature space, and interpretability is pursued through attention-map visualization. Quantization is discussed briefly to bolster deployment efficiency, and mixed precision for training efficiency. But, *the core experimental question is on the extent of context effects and size (scaling laws), on prediction accuracy* for a given set of data, well-engineered features, and a given attention style. The findings provide evidence on whether transformer-style architectures, built from first principles for financial data, meaningfully enhance predictive modeling and information capture in futures markets, as they do in natural language processing.

Methodology

The dataset comprises one continuous year of Nasdaq-100 futures trading data, transformed into custom-duration bars that aggregate tick activity into economically meaningful intervals. Each bar is labeled with one of three possible outcomes: upward, downward, or neutral, using a triple-barrier scheme calibrated to realistic intraday volatility thresholds. The preprocessing pipeline includes feature engineering, with the derived features to be described in detail, and rolling normalization that ensures comparability across regimes while preserving causal order. No pretrained weights or external initialization are used; each model is trained from scratch via supervised learning, mapping past bar sequences to next-bar class labels.

The experimental design compares three architectures: (1) six variants of a custom causal transformer trained autoregressively on the normalized feature sequences, (2) an ARIMA model representing a classical econometric baseline, and (3) a one-dimensional convolutional neural network as a deep learning benchmark. For transformers in autoregressive mode, the model *predicts only the next item in a sequence* and then (optionally) uses that prediction as input for the next step, but we will not here, based on the assumption that the model needs fresh data at every inference. This is a difference with the output of today's LLMs that I deem inconsequential to the threshold question of next token prediction. Training and evaluation are conducted on strictly forward-chained temporal splits to prevent leakage. The transformer models vary in size from 130M to 300M and 500M parameters, each a context length of 1024 and 2048. All models share the same input data and label structure to isolate architectural effects from data differences. The method is that we are testing transformers variants here, not tokenization approach or machine learning process.

ALiBi positional bias is employed to preserve temporal continuity and allow for extrapolation beyond the maximum trained context length. Causal masking enforces the unidirectional flow of information, ensuring that the model conditions only on past bars. Flash Attention accelerates training and inference while enabling longer context windows without prohibitive memory consumption. These design decisions are not treated as ablation variables but rather as theoretically necessary alignments between the architecture and the financial forecasting problem. They will be defined. Model training proceeds under a standard cross-entropy loss, possibly with balanced sampling to address class distributional imbalances.

This study adopts an autoregressive supervision paradigm because the nature of the financial forecasting problem is inherently sequential and causal: the next market state depends on the evolution of prior states, not on future information. Classic technical analysis in securities is founded on the assumption that estimations of the future are encapsulated in historical market action. In this framework, the transformer is trained on ordered sequences of bar-level data *for just the next bar*: each sequence representing a continuous stream of market activity, and the model learns to predict the next-bar label (UP, DOWN, or NONE) conditioned exclusively on preceding observations. The causal mask mirrors the temporal asymmetry that governs real-world trading decisions. Unlike disjoint sample-based learning, where each observation is treated independently, autoregressive supervision captures the dynamic dependencies and pathwise context that drive price formation and volatility clustering in financial markets. The approach remains fully supervised because each training step pairs an input sequence with a known ground-truth label and optimizes a cross-entropy loss, enabling the model to learn a mapping from historical features to the next market outcome in a statistically principled way. By encoding this causal structure directly into the learning architecture, the model's inductive bias matches the domain's information constraints, making autoregressive supervision not only methodologically appropriate but theoretically necessary for a realistic and disciplined study of predictive modeling in futures markets.

Supporting Analysis

Financial data are inherently non-stationary, with evolving volatility, liquidity, and structural regimes. To mitigate these effects, feature normalization and model retraining occur in rolling windows. This dynamic updating ensures that the parameters remain responsive to shifting market dynamics without introducing look-ahead bias. Complementing this, a brief entropy analysis quantifies the information content of the engineered features, serving to validate the complexity of the chosen representation. Entropy measures are computed across time to examine the richness and stability of predictive signals within the feature space. A proof of concept has already shown I have good features. Economic value is treated as an implicit derivative of predictive skill rather than an explicit target variable, distinguishing the predictive problem from downstream trading decisions. The barriers are symmetrical, so the risk/reward is 1:1, barring slippage. Evaluation is therefore grounded in directional accuracy, precision, and recall rather than portfolio simulation. This focus isolates the predictive capacity of the model architecture from the stochastic nature of realized returns.

Complementary Analyses

A key complementary analysis extends the interpretability and practicality of the results. First, transformer attention maps are visualized to determine whether the model allocates attention in economically plausible ways...for example, to volatility clusters, high-volume transitions, or feature interactions that resemble human chart analysis. These visualizations serve as both interpretive diagnostics and empirical checks on the model's internal logic. While broader topics such as transfer learning, diffusion forecasting, and multimodal data fusion are relevant in the literature, they remain outside the experimental scope of this study. They are acknowledged in the related work as possible extensions for future research rather than as confounding factors within the present design.

Challenges and Limitations

Predicting futures movements remains subject to well-known challenges: non-stationarity, high noise-to-signal ratios, and potential overfitting. Regularization and rigorous temporal validation are employed to counter these issues. The computational cost of training long-context transformers is substantial; Flash Attention and mixed precision training mitigate but do not eliminate this constraint and I am training and testing on a single consumer-grade GPU. I may run out of time to investigate all that is proposed here.

The triple-barrier labeling methodology introduces assumptions about barrier width and horizon; these assumptions are explicitly defended through observed descriptive statistics rather than sensitivity testing. For instance, in my data set, roughly 50% of bar range is above 6 points, and 50% below. This is chosen as the width of the barriers cleverly to focus on a spot where the presence of a trading opportunity is itself a random variable that need not be measured. This leverages a null hypothesis that the market is a random walk; if prediction can happen here then the past contains profitable information about the future.

Finally, the literature has begun to debate whether transformers outperform simpler architectures in time-series forecasting: an open question this study empirically addresses through direct comparison with ARIMA (autoregression) and CNN (fully supervised) baselines. Unlike the transformer's autoregressive supervision, the CNN does not model long-range dependencies across hundreds or thousands of bars; its receptive field is constrained by kernel size and depth, favoring localized temporal correlations over global context. Nor does it employ causal masking—the CNN relies on architectural design (non-centered convolutions and proper padding) to ensure that no future information leaks into the past. Conceptually, this makes the CNN a feed-forward, sliding-window predictor rather than a generative sequence model. It learns a deterministic mapping from a fixed historical horizon to the next outcome, reflecting a stationary approximation of the underlying process. The contrast here will be compared, mostly in performance.

Conclusion

By framing futures bar prediction as a causal sequence modeling task, this research examines whether transformer architectures built from scratch under supervised learning can discover predictive structure in complex financial data. The empirical evidence will illuminate the extent to which these attention-based mechanisms and efficient training techniques translate to meaningful gains in directional accuracy relative to classical and convolutional baselines. Studies have proposed improvements made to the classic transformer, like flash attention, and some of those will be adopted and exposed, but not tested. The reader will be directed towards other studies that prove their value. The investigation is performance and interpretability through attention at differing sizes and contexts. This contributes to the emerging dialogue in financial modeling on whether deep causal sequence architectures can provide robust, data-driven representations of market dynamics in the absence of explicit domain priors, but instead via attention. This is the primary conclusion presented.

An interesting thought: abstractly, my hypothesis formed is predicated on the postulate that thought-patterns in language and economics share a common ramification of a deeply conscious form. Although it will not be discussed expressly, there is a root of philosophy here from the great economist Frank Ramsey (February 1903 – January 19, 1930) who was one of the first scholars to paint a subjective picture of probability, that beliefs are in essence some bet. He further posited that rationality in all forms is “a kind of pragmatism: we judge mental habits by whether they work”, and in semantics generally, that pragmatism broadly construed is that “the meaning of a sentence is to be determined by reference to the actions in which asserting it would lead”. Thus the experimental idea tested here is founded in the postulate that linguistic patterns and financial behavior are in fact of the same *grounding* to an LLM, ergo transformers model a common grammar in NLP and financial times series. That postulate can’t be explored fully within the time we are given, but I will suggest that an inductive conclusion should be that utility for prediction is prima facie evidence for LLM grounding, because “if it don’t make dollars, it don’t make sense.”