Transformers and attention-based networks in quantitative trading: a comprehensive survey

Lucas Coelho e Silva
Aeronautics Institute of Technology
Brazil
coelho@ita.br

Gustavo de Freitas Fonseca
Aeronautics Institute of Technology
Brazil
gustavo.fonseca@ga.ita.br

Paulo Andre L. Castro
Aeronautics Institute of Technology
Brazil
pauloac@ita.br

Abstract

Since the advent of the transformer neural network architecture, there has been a rapid adoption and investigation of its applicability in various domains, such as computer vision, speech processing, and natural language processing, with the latter most notably exemplified by the rise of Large Language Models. These accomplishments have also led to increased interest in other network architectures that rely on attention mechanisms, one of the building blocks of transformers. Transformers and other attention-based networks are being applied to the quantitative analysis, management, and trading of financial assets, be it for price movement prediction, discovery of trading strategies, portfolio optimization, and risk management. The applications range across different asset categories, including equity markets, foreign exchange pairs, cryptocurrencies, and futures markets. This survey aims to provide a comprehensive overview of the applications of attention-based networks within the field of quantitative analysis, management, and trading of financial assets. After a brief overview of transformers and attention mechanisms, we analyze the existing applications of these architectures for quantitative finance in a taxonomy of four specializations: Alpha Seeking, Risk Management, Portfolio Construction, and Execution. After comparing the literature in light of the research problems, modeling approaches, and complementary results, we discuss current challenges and research opportunities.

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Applied computing**; • **General and reference** → **Surveys and overviews**;

Keywords

Quantitative trading, Machine Learning, Transformers

ACM Reference Format:
Lucas Coelho e Silva, Gustavo de Freitas Fonseca, and Paulo Andre L. Castro. 2024. Transformers and attention-based networks in quantitative trading: a comprehensive survey. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698684

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698684

1 Introduction

Artificial Intelligence (AI) has impacted several areas, with significant effects in finance, where various machine learning (ML) techniques are being employed to tackle complex challenges. More recently, the field of quantitative finance has been invested in novel research opportunities created by the introduction of attention-based neural networks. Among such architectures are transformers, a deep-learning neural network architecture that led to massive impacts in Natural Language Processing (NLP) [47], fueling the rise of Large Language Models (LLM) [6, 10, 45]. Building on attention mechanisms, transformers capture sequential patterns and long-range dependencies with no convolutions nor recurrence arrangements while being notably parallelizable [46]. Since its advent, there has been a rapid adoption and investigation of its applicability in many domains, such as computer vision [30] and speech processing [20]. These accomplishments have also led to increased interest in other neural network architectures that rely on attention mechanisms, one of the building blocks of transformers. The growing number of research initiatives on the application of these attention-based neural network architectures arise in various disciplines of the quantitative analysis, management, and trading of financial assets field, be it for price movement prediction [25], high-frequency trading [3, 18], portfolio optimization [42], order placement and execution [1], or risk management [38]. The applications range across different asset categories, including equity markets [29, 50], foreign exchange (Forex) pairs [12], cryptocurrencies [19], and futures markets, such as the crude oil market [15].

Influenced by its novelty, current literature on quantitative analysis, management, and trading of financial assets has targeted exploring the usage of transformers and other attention-based networks rather than providing an outlook of its building blocks, typical use cases, challenges, and future research directions.

This survey aims to fill this gap by providing a comprehensive overview of the applications of these neural networks within quantitative trading, whether it relates to its fundamentals and underlying theoretical tools, such as time-series prediction tasks or through its direct specializations, such as market risk management tools or the creation and operationalization of trading signals and allocation models that target alpha generation.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of transformers and attention mechanisms, and discusses the characteristics of transformers that both justify the research interest and lead to the applications in quantitative finance, especially for tasks that deal with tradable financial assets. Section 3 presents an outlook of how transformers are being used in quantitative trading, inspecting current literature for

common aspects of research, modeling approaches and major challenges that lead to future research opportunities. Finally, Section 4 details the conclusion by summarizing current research and future possibilities.

2 Basics of transformers and attention-based networks

In this section, we briefly cover the basics of transformers and other attention-based neural networks. We intend to provide the intuition behind these networks and their building blocks before discussing the aspects that justify and ground research of their applications within the field of quantitative analysis, management, and trading of financial assets, i.e., what about these designs make them a viable alternative for the field, and which of their features have sparked interest for research in this domain.

The development of transformers, proposed by [46], stems from the field of neural machine translation. It is a subfield of statistical machine translation (SMT) that uses neural networks for translating text from one language to another, with a single large model taking a sentence in one language and outputting the same sentence in another.

A core development in this field that ultimately led to the formulation of transformers is the introduction of the attention mechanism by [2]. It was designed to tackle the information bottleneck challenge within Encoder-Decoder Recurrent Neural Networks (RNN) [9] for sequence-to-sequence modeling [43]. Because the decoder of an RNN only has access to the final state of the encoder, the encoder has to represent the meaning of the entire input sequence. Attention is precisely the mechanism that allows the decoder to access all of the encoder's hidden states. Nevertheless, relying simultaneously on all states would drastically increase the dimensionality of the decoder inputs. Thus, the design of attention includes a weighing mechanism to define which encoder states to prioritize or "attend". This enables the attention-based models to learn to focus on the most essential parts of a sequence for each instant, creating ways to represent various alignment arrangements.

Transformers keep the Encoder-Decoder approach and extend the concept of attention, relying on a self-attention design. It abdicates from recurrences and convolutions [46], allowing for large sequences to be used within a weighted averaging mechanism that generates attention scores for each sequence component. Particularly, transformers implement the self-attention layer via the scaled dot-product attention, a formulation that computes the attention scores by using dot products as a similarity function, including a scaling factor to prevent large numbers. By dropping the recurrence, however, transformers demanded a new mechanism for representing the relative positions between the sequence points, a problem solved by adopting positional encodings [46]. Lastly, the self-attention mechanism is implemented with multiple attention heads, which, in practice, allows for the model to focus on several aspects of similarity simultaneously.

The clear modularization that leads to the separation of concerns within the design of transformers promotes research that tackles modifying the architecture for different tasks. For instance, the literature and current practice have also explored decoder-only [37] and encoder-only [10] designs. These are particularly suitable for autoregressive behavior and conversion of inputs into meaningful numerical representations, respectively. Furthermore, architecture modifications might target different goals, such as reducing complexity and adding priors to the attention mechanism [17], or providing variants that target specific data structures and patterns, such as improving the efficiency of Transformers for long sequences [23] or better representing autocorrelation phenomena and seasonality in long-term time series forecasting [49, 54, 55]. This ultimately originated several transformer-inspired networks, such as Gaussian Transformers [14], and the Reformer [23], Informer [54], Autoformer [49], and FEDformer [55] architectures.

In finance, transformers have been driving research interest due to a few factors. First, as they were originally designed to handle text sequences, it is straightforward to apply them to other types of sequential data, including time series data, which is prevalent in the field. Second, their attention mechanisms can potentially identify relevant market factors, including phenomena at different timescales, while their scalability allows for analyzing large amounts of market data. Finally, the modularized design adds flexibility that allows for exploring novel architectures tailored to suit the needs of quantitative analysis, management, and trading of financial assets.

3 Transformers in quantitative trading

In order to fulfill our objective of providing a comprehensive overview of the applications of transformers and other attention-based neural networks for quantitative trading, we select references that thoroughly represent the possibilities of research, applications, and challenges within the field. We analyze current literature related to both the fundamental disciplines that compose the field, as well as to its direct specializations that deal with the creation and operationalization of trading signals and allocation models that target alpha generation or applications for managing and minimizing market risk. We are thus interested in any phase of a development process whose final stage enables the interface between a system and a market via a set of algorithms without human discretionary intervention, including every aspect that influences the decision of which market and instrument to choose, which market action to take and when, until which condition to hold a decision, the criteria that lead to the decision, and so forth. Therefore, we include in our analysis a broad number of topics that range from the characterization of market data – such as time series of price data, which could drive the discovery of trading signals – to order execution mechanisms particular to a given market.

We examine the existing applications of transformers and other attention-based networks from the perspective of specializations within the field of quantitative trading. Namely, we focus on the four specializations described by [33]: Alpha Seeking, Risk Management, Portfolio Construction, and Execution. This is complemented by the analysis of research regarding its discipline and core tasks, such as financial time series forecasting, market point prediction, characterization, and others. Finally, it is noteworthy that the specializations and disciplines are often self-interacting and complementary approaches rather than competing and non-intersecting ones, making the delineation between these fields nuanced and less distinct in certain applications.

3.1 Alpha Seeking

Central to general quantitative finance and with results that directly impact the developments of Portfolio Construction, Risk Management, and Execution, the specialization of Alpha Seeking provides a comprehensive set of components and techniques that serve as a backbone of the field of quantitative trading both in terms of modeling as well as the construction real-world systems that interact with financial markets. It ranges from fundamental investigations, such as forecasting market time series, to applied techniques that focus on evaluating a model's outputs within market scenarios via techniques such as backtesting. The applications include various asset classes and trading regimes, such as Low-Frequency Trading (LTF) and High-Frequency Trading (HFT), and investigate multiple data sources, such as price data, fundamental data, and sentiment data from text sources.

In this sense, the opportunities for exploring statistical learning techniques are manifold, and novel ML designs promptly trigger research interest within the specialization. This is precisely the case with transformers and other attention-based networks, with recent literature focusing on their applicability for the direct development of trading signals and algorithms or characterization and prediction of future market states. Recent investigations on Alpha Seeking examine topics such as the suitability of these architectures for price movement prediction and mining of trading signals, including different sources and arrangements of data; effects of ensembles of attention-based models; extraction of features for financial modeling; impacts of model specialization, both in terms of data sources as well as different aspects of a quantitative trading problem; and novel attention-based architectures tailored for the specific tasks related to constructing an alpha model.

On the development of novel transformer-based architectures, for instance, [3] investigates the application of different attention-based designs for forecasting the log-return of Bitcoin within an HFT trading strategy application that relies on Limit Order Book (LOB) data. In addition to the transformer, they investigate the Autoformer [49] and FEDformer [55] variants along with a proposed HFformer architecture that performs quantile regression of the log returns while featuring time-dependent spiking activation [4] functions with learnable thresholds. While there are only two days of out-of-sample backtesting, HFformer achieved higher cumulative profit-and-loss (PnL) results when compared with the Long short-term memory (LSTM) baseline while also outperforming the other transformer-based architectures.

A comprehensive group of attention-based networks is also explored by [5] for tasks of forecasting the LOB mid-price, the LOB mid-price difference, and the LOB mid-price movement. They define an LSTM-based architecture as the benchmark and compare it to the performance of transformers, Autoformers [49], Informers [54], Reformers [23], and FEDformers [55]. The study finds the LSTM-based network to yield better and more robust predictions regarding price differences and price movements, with the transformer-inspired models outperforming during the absolute-price sequence prediction task. Nevertheless, the attention-based architectures also displayed profitable results under the study's experimental setup.

Generating signals for HFT in Forex markets is the focus of [18]. For that, they adapt the canonical transformer architecture by removing the decoder while focusing on representing the market data on different time scales. While it lacks the description of the naive strategy used as a benchmark, the research suggests transformers to be a promising technique for the discovery of profitable trading strategies.

The models can also be constructed with different feature space dimensions. [11] investigates the application of a transformer with time embeddings for Forex price movement prediction while also assessing whether multivariate features from fundamental and technical analysis improve performance. After backtesting the model predictions, the study suggests performance improvements with the multivariate features for the transformer architecture, which outperformed the benchmark. Similarly, [12] further explores applications of transformers for Forex trading and focuses on intraday operations to investigate whether transformer-based networks exhibit predictive capabilities in this market. According to the reported results, the transformer-based approach yielded better performance than the ResNet-LSTM benchmark.

Alternative to the definition of LSTMs purely as benchmarking tools, the literature also investigates the integration between LSTMs and attention-based networks. For instance, [29] investigates the forecasting of market trends within a Non-stationary Markov Chain (NMC) approach with a combination of Bidirectional Encoder Representations from Transformers (BERT) and LSTM networks. The designed architecture achieved satisfactory metrics in terms of forecasting accuracy and annualized returns of the trading signals.

Whether combining multiple models is beneficial is also a research question that sparks interest. [35] investigates ensemble mechanisms of temporal transformers trained with sliding windows for price forecasting in equity markets. The study first points out the inability of canonical transformers to capture the market volatility in local contexts. To counteract this shortcoming, they propose temporal transformers with similarity embeddings for improving performance in multi-horizon forecasts. While the temporal transformer models present over 40% in performance improvements when compared to the base one, they are frequently unable to fully utilize the time series data. The results also suggest that the ensemble model leads to performance improvements, as it potentially uses learners focused on heavily disjoint periods.

The different ways to arrange price data, the integration of multiple data sources, and the variety of time frames also lead to investigations. For instance, [15] apply the vanilla transformer architecture for price prediction of Shanghai crude oil futures based on Open-High-Low-Close (OHLC) price data. The achieved results of up to 30% increase in performance metrics, along with the attained backtesting results, suggest the transformer model presents out-of-sample forecasting capability in various time frames, ranging from 30 minutes to monthly frequencies.

Regarding secondary data sources that go beyond direct market data such as prices and volumes, [25] propose a so-called transformer-encoder-attention network architecture for extraction of financial features from social media text and use it in combination with stock prices for the ultimate goal of price movement prediction via a classifier in equity markets. While the results were not

validated in terms of backtests but rather statistical learning metrics such as accuracy and Matthew's Correlation Coefficient, they suggest the superior performance of the attention-based networks when compared with benchmarks such as autoregressive and LSTM models.

There are multiple approaches to sentiment and text analysis, however. A different approach from that of [25] is to directly apply LLMs for processing text, such as news data, as a way to derive novel features and trading signals. LLMs are being thoroughly explored for finance applications, and [26] presents a practical survey focused on a broader view of the topic. Here, we discuss solely the direct implications for quantitative trading.

For instance, [31] addresses whether Generative Pre-trained Transformer (GPT)-based and other LLMs can analyze news of a particular firm and assess if the reported content is neutral, positive, or negative for the company stock prices, all done via prompts. They compare the recommendations with the next day stock returns and find positive correlations between GPT-3 and GPT-4 analyses with the following returns, especially for stocks with smaller market capitalization values, while simpler models such as GPT-1, GPT-2, and BERT do not possess predictive power. The results indicate that state-of-the-art LLMs are a promising alternative to traditional methods of sentiment analysis. However, validation is particularly controversial when using pre-trained proprietary models. Even though researchers rely on the fact that LLMs were trained with information up to a point in time to define the cut-off dates for the out-of-sample data, the fine-tuning processes constantly happening beyond the pre-training dates can introduce new information and lead to data leakage.

These same predictive behaviors arise in cryptocurrency markets, as shown by [19], with the difference that even simpler models like BERT contribute to positive performance metrics when applied within a Bitcoin trading system that incorporates the sentiment-based extracted features.

Another topic of research is investigating if constructing specialized models for different languages yields more precise and informative sentiment analysis. In this scope, [52] focuses on deriving a technical indicator that measures the level of desirability to invest in Chinese company stock. They compare the results generated by baseline LLMs with Erlangshen, a model trained on Chinese corpora. The results indicate that, despite the order of magnitude difference in the number of parameters of the much smaller Erlangshen model, the application of pre-training and fine-tuning techniques specific to the local language yields better results for the extracted trading signal.

Alternatively, whether a model should specialize in simultaneously learning multiple aspects of the problem is also a topic of research interest. [48] focuses on constructing a model that can learn both the position side and size for trading futures contracts based on price series solely. It does so by directly optimizing the Shape ratio during model training. The proposed Momentum Transformer model builds on the concept of Temporal Fusion Transformers (TFT), a hybrid architecture that, like [29], combines LSTMs and self-attention-mechanism capabilities. In addition to outperforming the pure LSTM-based model used as the benchmark, the proposed Momentum Transformer is interpretable and provides insights into trading strategies.

A different learning approach to modeling both trade directions and optimum sizing mechanisms is reinforcement learning (RL). [50] utilizes both transformer-based and U-Net neural networks within a deep RL context with the proposal of an end-to-end model for single stock trading named DRL-UTrans. The model takes as inputs windowed stock price sequences and outputs both the trading action and the weight of the action. The study indicates that DRL-UTrans outperforms the baseline approaches while also suggesting its effectiveness when facing market volatility and crashes.

3.2 Risk Management

Modeling risk is crucial to the field of quantitative analysis, management, and trading of financial assets. This is demonstrated by the multiple approaches that inherently take risk factors into consideration within all of the other three defined specializations. Nevertheless, Risk Management provides its own applications for a broader identification, assessment, and mitigation of potential risks associated with financial markets, thus deserving dedicated attention.

Statistical learning applications are particularly useful for tasks such as the prediction of market volatility and tail risks. When using attention-based networks, the application of these designs could be within hybrid approaches, combining the usage of neural networks with canonical methods, or within self-contained, attention-first applications. For instance, [38] targets the problem of creating accurate models for stock-index volatility that yield appropriate risk measures of equity markets. This application combines several autoregressive models and LSTM units with both the vanilla transformer and the Multi-Transformer, with the latter being an ensemble-based variant of the base architecture. The results suggest that, while the hybrid models based on multi-transformers heavily rely on regularization during the training process, they lead to more accurate risk measures when compared to other autoregressive algorithms and modeling approaches based on LSTMs or feed-forward networks. [41] evaluates a comprehensive group of stocks and attention-based networks to forecast realized volatility, addressing the claims made by [51] that attention-based models show limited capability of handling time-indexed data. It explores different forecast horizons (one business day, one business week, and two business weeks ahead forecasts) with the application of TFT [28], Informer [54], Autoformer [49], and PatchTST [34] networks. The results, addressed in light of the comparisons against benchmarks that include Neural Basis Expansion Analysis for Interpretable Time Series with exogenous variables (N-BEATSx), Neural Hierarchical Interpolation for Time Series (NHITS), and Heterogeneous Autoregressive (HAR) formulations, suggest that first-generation transformers, such as TSTs, underperform in financial forecasting. Nevertheless, the second-generation models of Informers, Autoformers, and PatchTSTs display efficient results even in scenarios with limited historical data. For instance, the investigated Autoformer outperformed in short-term forecasting tasks, and the results indicate the benefits of nuanced applications of these networks, tailored to specific contexts of data availability and forecast horizons.

The prediction and characterization of volatility also play an essential role in the market of options, with implications for pricing

these derivatives as well as the development of hedging strategies. This is the focus of the approach proposed by [21], which presents five different neural-network architecture designs to predict the volatility surfaces based on S&P 500 options data. The research explores the application of standard physics-informed neural networks, convolutional long-short-term memory (ConvLSTM) networks, self-attention ConvLSTMs, and a so-called physics-informed convolutional transformer. By focusing on a novel physics-informed network design, the solution process benefits from taking into consideration the governing differential equations and necessary boundary conditions directly on the loss function to be minimized. While the five different architectures are able to successfully predict volatility surfaces, the novel architecture based on convolutional layers and transformers outperforms the other formulations.

Finally, [44] tackles the research problem of the effective hedging of derivatives via SigFormer, a modeling approach that combines path signatures with transformers, as a way to effectively capture and represent complex patterns as well as extracting features from those sequences. SigFormer presents a faster learning process and robustness on simulated environments with synthetic data and displays positive results on backtests with S&P 500 index hedging.

3.3 Portfolio Construction

Portfolio Construction is one of the core and most thoroughly researched specializations in the field of quantitative analysis, management, and trading of financial assets. By focusing on the identification of efficient mixtures of alpha models and rebalancing mechanisms through time, this specialization creates opportunities for the development of models with multiple responsibilities and closely interfaces with the developments of Section 3.1, especially regarding tasks on predicting future asset performance. Because the optimum mixtures heavily rely on past performance and forecasts of future behavior, the application of neural networks with self-attention mechanisms has driven research interest. Recent research encompasses applications of these transformer-inspired networks for tasks that include the ranking of financial instruments; investigation of different attention formulations and their impact on portfolio construction; direct learning of optimum portfolio weights with custom loss functions; and deep RL-based approaches for estimating policies for constructing portfolios.

For instance, [39] proposes the Rank Transformer, a model that follows the Learning to Rank (LTR) approach to portfolio construction. It closely follows the canonical Transformer implementation, with minor adjustments on activation functions and normalization techniques. It incorporates the ranking objectives directly on the training process by optimizing the Normalized Discounted Cumulative Gain, in addition to returns, and delivers superior Shape Ratio and annualized returns than baseline methods such as the minimum variance and hierarchical risk parity strategies for both equity and Forex markets.

On another approach, [53] relies on time series market data of Chinese stocks for building a model that predicts stock returns within a portfolio construction context. They apply transfer learning from sentiment analysis as an attempt to capture long-range dependencies on the data. In terms of network architecture, [53] replaces the word embedding layer of the canonical transformer with a standard linear layer while also simplifying the decoder with the removal of autoregressive properties. In this modeling approach, the ranking of the predicted returns yields the stock portfolio, and the transformer-based approach yielded superior performance in predicting trends when compared to benchmarks while also delivering satisfactory risk-adjusted performance metrics.

The emphasis on risk-adjusted performance metrics within portfolios is also a goal of [42], as it introduces the direct adoption of risk-averaged returns as the objective function of the neural network training process. The network is designed to directly output the weight of each stock, from -1 (maximum short position) to 1 (maximum long position), and it is composed of convolutions and graph attention mechanisms, with the former being responsible for the direct processing of the raw financial features. The chosen model outperformed the selected benchmarks, both in terms of financial portfolio metrics and statistical learning performance metrics.

Similarly focusing on risk-adjusted performance metrics within portfolios, [7] investigates different formulations of the attention mechanism within network architectures trained with a custom Sharpe Ratio loss function for composing portfolios. The study pairs additive attention and self-attention mechanisms with LSTMs and RNNs with Gated Recurrent Units (GRU), compared against the RestNet architecture, as well as vanilla LSTMs and GRUs, with results indicating better performance by the models with additive attention paired with GRUs.

The Portfolio Construction specialization is also comprehensive in terms of the scope of assets, ranging from equities to cryptocurrencies. In particular to the latter, [40] proposes a method named the N-BEATS Perceiver, an attention-based design constructed on top of the Perceiver IO architecture [16], applied within a task of univariate time series point forecasting of cryptocurrency prices, closely related to the developments of Section 3.1, but also with portfolio construction applications. The proposed model outperforms the other investigated approaches in terms of portfolio risk profile and forecast accuracy. The study also provides insights into the scalability of transformer-based models by mentioning that the N-BEATS transformer variant presented sharp increases in memory during operations, in contrast to the better-performing N-BEATS Perceiver.

Finally, as in Alpha Seeking, RL is an alternative learning approach to modeling the portfolio construction problem. [22] explores a deep RL actor-critic formulation with the proposition of a variation of the transformer architecture that implements two-dimensional, relative position multi-head attention with gating layers. The proposed workflow relies on the transformer variant for both actor and critic, including the target function approximators. The portfolio performance analysis includes costs, with the proposed design outperforming alternative approaches such as Markowitz's Mean-Variance model and a uniform constant rebalanced portfolio.

3.4 Execution

While the discovery of trading signals and profitable trading strategies is of cornerstone relevance to the field of quantitative analysis, management, and trading of financial assets, being able to ensure

that the strategies are accurately reproduced in real-world applications with efficient market operations is fundamental. This gives rise to the specialization of Execution, which deals with core disciplines such as order management and trade processing. Relevant statistical learning tasks involve predicting and modeling order book dynamics and market liquidity, as well as the investigation of different order placement strategies.

One of these tasks refers to the problem of estimating the time-to-fill – i.e., the amount of time it takes to fill a limit order placed on the order book. [1] estimates the time-to-fill of orders placed in different book levels on Nasdaq. They investigate the usage of attention-based networks within a survival analysis framework and find that the convolutional attention network outperforms the state-of-the-art benchmark. Nevertheless, the results were not validated in terms of direct market application and simulation, leaving opportunities for future research.

A different problem within the specialization is modeling the order book dynamics in financial market simulations (FMS). The task is non-trivial, especially regarding the calibration of the generative process in a way that provides a good representation of the observed data in actual markets. For that, [27] focuses on learning vectorized representations of the LOB for calibrating a Preis-Golke-Paul-Schneid (PGPS) FMS model. It does so with an autoencoder neural network design with stacked transformer blocks for capturing temporal correlations and reconstructing the data, finding the proposed model to outperform in terms of reconstruction errors against other neural network approaches such as LSTMs and Convolutional Neural Networks (CNN). [36] carries out a similar approach to learning vectorized LOB representations via a transformer autoencoder but with different end goals. Instead, it targets modeling normal behavior for anomaly detection within the LOB, achieving state-of-the-art performance.

Finally, optimal execution and operationalization tasks are notably close to market-making (MM) initiatives. For instance, [13] relies on pre-trained networks with convolutions and transformers layers for LOB feature extraction within MM. The approach outperformed the baselines on a simulated environment while also providing explanatory insights on the dynamics of MM agents.

3.5 Summary of findings, common aspects of research and open challenges

The choice to define and categorize the research initiatives in the four specializations of Alpha Seeking, Risk Management, Portfolio Construction, and Execution provides us with the building blocks for finally establishing a comprehensive view of the application of attention-based neural networks for quantitative trading and map convergent ideas and techniques that may guide future developments that lead to new advancements.

Despite the silos, this section's examination of current literature and recent developments confirms the self-interacting complementary nature of the approaches, both in terms of challenges as well as techniques. The surveyed research spans a wide range of topics in the application of transformers and other attention-based architectures for quantitative trading. Investigated problems encompass tasks such as assessing the suitability of these attentive designs for financial data, including the integration of multiple data sources; predicting, simulating, and characterizing market and asset dynamics in terms of price, returns, volatility, and order execution; discovering and engineering relevant features for financial modeling; and weighting and ranking different alpha models when composing portfolios. Recurring approaches involve jointly and efficiently learning network representations for each task, be it via custom loss functions that integrate financial objectives such as risk-adjusted performance of trading signals and portfolios, or within an RL framework that directly assesses the impact of decisions for learning suitable policies and value functions. Table 1 summarizes this outlook and presents an overview of the literature selected to thoroughly represent the possibilities of research, applications, and challenges within the field of quantitative analysis, management, and trading of financial assets.

Overall, the results reported in the literature show promising ways regarding the applicability of transformers and other attention-based networks within the field, as they frequently surpass the alternative models used as benchmarks. More specifically, the successful approaches of distinct network architectures that incorporate self-attention mechanisms tailored to specific problems push for research on new architecture variations and loss functions that inherently carry information about the market and trading environment. The attentive networks have also been successfully used within hybrid contexts, deeming straightforward future developments that combine the networks with other techniques. In addition, the literature would benefit from future research on foundational work focused on understanding the reasons and contexts in which attention-based network variants work best in quantitative trading and mapping which components within these designs are relevant for achieving target results with different markets and data regimes.

Results are not always in agreement, and the disaccord also creates opportunities for further investigations. For instance, when forecasting the realized volatility of stocks, [41] found first-generation Transformer models, such as TFTs, to underperform in financial forecasting. On the other hand, [48] found the approach to outperform the benchmarks when used within a hybrid architecture and directly optimizing financial metrics during model training. Furthermore, that lack of diverse benchmarks, with comparisons mostly against LSTM models, prompts comparisons between the novel approaches and other machine learning techniques that, for instance, rely on established tabular data techniques.

Other research opportunities arise as recurrent topics from the analyzed initiatives, including the exploration of ensemble techniques, hybrid approaches [31, 41], transfer learning [12], online, continuous learning, and RL-based formulations [5, 35, 44, 48], and the integration of different data sources [12, 15]. On the other hand, other topics remain largely underexplored, such as the efficiency and costs of the usage of transformers in practical settings of quantitative finance and trading. While addressed by [5], this topic remains a gap worth exploring.

Some topics remain fully unexplored within the domain of transformers for quantitative trading. For example, recent developments within the general literature of transformers investigate uncertainty quantification frameworks such as Conformal Prediction coupled with the network design process [24], a line of research that directly impacts finance.

Table 1: Summary of literature of transformers and attention-based networks for quantitative trading.

Ref.	Specialization	Research problem	Modeling approach	Market
[3]	Alpha	Predicting log-return of Bitcoin with LOB data.	Transformer, Autoformer, FEDformer, and HFformer architectures.	Crypto
[5]	Alpha	Prediction of LOB mid-price, LOB mid-price difference, and LOB mid-price movement.	Transformer, Autoformer, Informer, Reformer, and FEDformer architectures, compared against LSTMs.	Crypto
[18]	Alpha	Modeling signals for high-frequency trading.	Vanilla Transformer without the decoder, coupled with an Exponential Moving Average model applied to the input for representing data on different time scales.	Forex
[11]	Alpha	Price movement prediction.	Transformer with time embeddings for regression and classification tasks.	Forex
[12]	Alpha	Price movement prediction.	Transformers trained on top of price and technical analysis features such as Bollinger bands and Relative Strength Index. Benchmarked against ResNet LSTMs and a ResNet network.	Forex
[29]	Alpha	Price forecasting that yield a trading signal based on the movement.	Hybrid approach with BERT and LSTMs.	Futures
[35]	Alpha	Price forecasting.	Ensemble of sliding-window temporal transformers.	Equity
[15]	Alpha	Price forecasting.	Canonical transformer.	Futures
[25]	Alpha	Price movement prediction.	Transformer Encoder Attention.	Equity
[31]	Alpha	Sentiment analysis for creating features and trading signals.	DistilBart-MNLI-12-1, GPT-1, GPT-2, GPT-3, GPT-4, BERT.	Equity
[52]	Alpha	Sentiment analysis for creating features and trading signals.	Chinese-GPT, Chinese-FinBERT, Erlangshen-RoBERTa110M-Sentiment.	Equity
[19]	Alpha	Sentiment analysis for creating features and trading signals.	Pre-training of crypto-specific language models based on BERT.	Crypto
[48]	Alpha	Learning of trading action and weights (position size).	Momentum Transformer, a hybrid architecture that combines LSTMs and self-attention-mechanism capabilities, directly optimizing the Shape Ratio during training.	Futures
[50]	Alpha	Learning of trading action and weights (position size).	Transformer-based and U-Net neural networks within a deep RL context for end-to-end models for single stock trading.	Equity & cryptos
[38]	Risk	Volatility forecasting.	Hybrid approach with Transformers, Multi-Transformers, autoregressive models, and LSTM units.	Equity
[41]	Risk	Volatility forecasting.	TFT, Informer, Autoformer, and PatchTST, compared against N-BEATSx, NHITS, and HAR.	Equity
[21]	Risk	Modeling and predicting volatility surfaces.	PINN, ConvLSTM, self-attention ConvLSTM, phyisics-informed convolutional transformer.	Options
[44]	Risk	Hedging derivatives	SigFormer: a novel architecture that couples transformers with path signatures.	Options
[39]	Portfolio	Construction of portfolios via ranking.	Rank Transformer, a model that allocates assets by predicting the rank of instrument efficiency.	Equity & Forex
[53]	Portfolio	Prediction of stock returns for portfolio composition.	Transfer learning from sentiment analysis applications. Adapted transformer architecture.	Equity
[42]	Portfolio	Directly output weights for each portfolio component.	Arrangement of convolutions and graph-attention mechanisms, with risk-averaged returns used during training.	Equity
[40]	Portfolio	Point forecasting of cryptocurrency prices and portfolio composition.	N-BEATS Perceiver, constructed on top of the Perceiver IO architecture.	Crypto
[7]	Portfolio	Directly output weights for each portfolio component.	Additive attention and self-attention mechanisms with LSTMs and GRUs.	Equity.
[22]	Portfolio	RL for portfolio construction.	Deep RL actor-critic formulation with a transformer variant that implements two-dimensional attention with gating layers.	Equity
[1]	Execution	Fill-time prediction of limit orders.	Combination of attention layers and convolutions.	Equity
[27]	Execution	Learning vectorized representations of LOB for market simulations.	Autoencoder neural network design with stacker transformer blocks for capturing temporal correlations.	Equity
[36]	Execution	Learning LOB representations for anomaly detection.	Autoencoder based on transformers for unsupervised capturing of temporal vector representations.	Equity
[13]	Execution	Modeling LOB for market making.	Combination of convolutions and attention for extracting features from the LOB.	Equity

Similarly, the quantitative trading literature can benefit from recent developments regarding the usage of transformers within Automated Machine Learning (AutoML) frameworks [32], including Neural Architecture Search (NAS) methods [8]. For instance, AutoML can be particularly useful for feature extraction and quick investigation of trading signals and strategies. Additionally, the NAS subset of techniques enables efficient processes for defining suitable architectures, including the creation of hardware-aware models with improved inference latency and performance [8], which are critical aspects of real trading systems.

Finally, a common challenge faced by the analyzed research regards validation. Most of the analyzed initiatives resort to train-test or train-validation-test splits, with single-set estimations of model performance. While this might indicate trends, especially when considering the collective positive results reported by most studies, literature would benefit from the development of novel methods that could drive the field towards standardized practices that further improve reproducibility aspects and hinder false discoveries of strategies and models.

4 Conclusions

This survey addresses the applications of transformers and other attention-based networks within the field of quantitative analysis, management, and trading of financial assets. We have analyzed the existing applications of these architectures for quantitative trading in a taxonomy of the specializations of Alpha Seeking, Portfolio Construction, Risk Management, and Execution [33]. After comparing the literature in light of the research problems, modeling approaches, and complementary results, we discuss current challenges that lead to future research opportunities.

The surveyed research covers a wide range of topics in the application of transformers and other attention-based architectures for quantitative trading. Investigated problems include tasks such as assessing the suitability of these attentive designs for financial data, including the integration of multiple data sources; predicting, simulating, and characterizing market and asset dynamics in terms of price, returns, volatility, and order execution; weighting and ranking different alpha models when composing portfolios; and discovering and engineering relevant features for financial modeling. Recurring approaches involve jointly and efficiently learning network representations for each task, be it via custom loss functions that integrate financial objectives such as risk-adjusted performance of trading signals and portfolios, or within Reinforcement Learning frameworks that directly assesses the impact of decisions for learning suitable policies and value functions.

There are several challenges in the field that pose promising directions for further research on the application of attention-based networks. These challenges include possibilities regarding further exploration of ensemble and transfer learning techniques, as well as working towards advancements in validation mechanisms that further improve reproducibility aspects and hinder false discoveries of strategies and models. Finally, some topics remain fully unexplored and provide multiple opportunities for future research. These include coupling the network design process with uncertainty quantification frameworks, such as Conformal Prediction, or the usage of transformers within AutoML frameworks.

References

[1] Álvaro Arroyo, Álvaro Cartea, Fernando Moreno-Pino, and Stefan Zohren. 2024. Deep attentive survival analysis in limit order books: estimating fill probabilities with convolutional-transformers. *Quantitative Finance* 24, 1 (Jan. 2024), 35–57. https://doi.org/10.1080/14697688.2023.2286351

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. http://arxiv.org/abs/1409.0473 arXiv:1409.0473 [cs, stat].

[3] Fazl Barez, Paul Bilokon, Arthur Gervais, and Nikita Lisitsyn. 2023. Exploring the Advantages of Transformers for High-Frequency Trading. http://arxiv.org/abs/2302.13850 arXiv:2302.13850 [cs, q-fin].

[4] Trevor Bekolay, James Bergstra, Eric Hunsberger, Travis DeWolf, Terrence C. Stewart, Daniel Rasmussen, Xuan Choo, Aaron Russell Voelker, and Chris Eliasmith. 2014. Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics* 7 (2014). https://doi.org/10.3389/fninf.2013.00048

[5] Paul Bilokon and Yitao Qiu. 2023. Transformers versus LSTMs for electronic trading. http://arxiv.org/abs/2309.11400 arXiv:2309.11400 [cs, econ, q-fin].

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165 arXiv:2005.14165 [cs].

[7] Hieu K. Cao, Han K. Cao, and Binh T. Nguyen. 2020. DELAFO: An Efficient Portfolio Optimization Using Deep Neural Networks. In *Advances in Knowledge Discovery and Data Mining*, Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan (Eds.). Vol. 12084. Springer International Publishing, Cham, 623–635. https://doi.org/10.1007/978-3-030-47426-3_48 Series Title: Lecture Notes in Computer Science.

[8] Krishna Teja Chitty-Venkata, Murali Emani, Venkatram Vishwanath, and Arun K. Somani. 2022. Neural Architecture Search for Transformers: A Survey. *IEEE Access* 10 (2022), 108374–108412. https://doi.org/10.1109/ACCESS.2022.3212767

[9] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805 arXiv:1810.04805 [cs].

[11] Tizian Fischer, Marius Sterling, and Stefan Lessmann. 2024. Fx-spot predictions with state-of-the-art transformer and time embeddings. *Expert Systems with Applications* 249 (Sept. 2024), 123538. https://doi.org/10.1016/j.eswa.2024.123538

[12] Przemysław Grądzki and Piotr Wójcik. 2024. Is attention all you need for intraday Forex trading? *Expert Systems* 41, 2 (Feb. 2024), e13317. https://doi.org/10.1111/exsy.13317

[13] Hong Guo, Jianwu Lin, and Fanlin Huang. 2023. Market Making with Deep Reinforcement Learning from Limit Order Books. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Gold Coast, Australia, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191123

[14] Maosheng Guo, Yu Zhang, and Ting Liu. 2019. Gaussian Transformer: A Lightweight Approach for Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 6489–6496. https://doi.org/10.1609/aaai.v33i01.33016489

[15] Wenyang Huang, Tianxiao Gao, Yun Hao, and Xiuqing Wang. 2023. Transformer-based forecasting for intraday trading in the Shanghai crude oil market: Analyzing open-high-low-close prices. *Energy Economics* 127 (Nov. 2023), 107106. https://doi.org/10.1016/j.eneco.2023.107106

[16] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. http://arxiv.org/abs/2107.14795 arXiv:2107.14795 [cs, eess].

[17] Uday Kamath, Kenneth L. Graham, and Wael Emara. 2022. *Transformers for Machine Learning: A Deep Dive* (1 ed.). Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781003170082

[18] Konstantinos T. Kantoutsis, Adamantia N. Mavrogianni, and Nikolaos P. Theodorakatos. 2024. Transformers in High-Frequency Trading. *Journal of Physics: Conference Series* 2701, 1 (Feb. 2024), 012134. https://doi.org/10.1088/1742-6596/2701/1/012134

[19] Gyeongmin Kim, Minsuk Kim, Byungchul Kim, and Heuiseok Lim. 2023. CBITS: Crypto BERT Incorporated Trading System. *IEEE Access* 11 (2023), 6912–6921. https://doi.org/10.1109/ACCESS.2023.3236032

[20] Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An Efficient Transformer for Automatic Speech Recognition. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 9361–9373. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3ccf6da39eeb8fefc8bbb1b0124adbd1-Abstract-Conference.html

[21] Soohan Kim, Seok-Bae Yun, Hyeong-Ohk Bae, Muhyun Lee, and Youngjoon Hong. 2024. Physics-informed convolutional transformer for predicting volatility surface. *Quantitative Finance* 24, 2 (Feb. 2024), 203–220. https://doi.org/10.1080/14697688.2023.2294799

[22] Tae Wan Kim and Matloob Khushi. 2020. Portfolio Optimization with 2D Relative-Attentional Gated Transformer. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, Gold Coast, Australia, 1–6. https://doi.org/10.1109/CSDE50874.2020.9411635

[23] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*. ICLR, Online, N/A. https://doi.org/10.48550/arXiv.2001.04451

[24] Junghwan Lee, Chen Xu, and Yao Xie. 2024. Transformer Conformal Prediction for Time Series. http://arxiv.org/abs/2406.05332 arXiv:2406.05332 [cs].

[25] Yawei Li, Shuqi Lv, Xinghua Liu, and Qiuyue Zhang. 2022. Incorporating Transformers and Attention Networks for Stock Movement Prediction. *Complexity* 2022 (Feb. 2022), 1–10. https://doi.org/10.1155/2022/7739087

[26] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *4th ACM International Conference on AI in Finance*. ACM, Brooklyn NY USA, 374–382. https://doi.org/10.1145/3604237.3626869

[27] Yuanzhe Li, Yue Wu, and Peng Yang. 2024. SimLOB: Learning Representations of Limited Order Book for Financial Market Simulation. http://arxiv.org/abs/2406.19396 arXiv:2406.19396 [cs].

[28] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[29] Chang Liu, Jie Yan, Feiyue Guo, and Min Guo. 2022. Forecasting the Market with Machine Learning Algorithms: An Application of NMC-BERT-LSTM-DQN-X Algorithm in Quantitative Trading. *ACM Transactions on Knowledge Discovery from Data* 16, 4 (Aug. 2022), 1–22. https://doi.org/10.1145/3488378

[30] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. 2024. A Survey of Visual Transformers. *IEEE Transactions on Neural Networks and Learning Systems* 35, 6 (June 2024), 7478–7498. https://doi.org/10.1109/TNNLS.2022.3227717

[31] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. http://arxiv.org/abs/2304.07619 arXiv:2304.07619 [cs, q-fin].

[32] Ambarish Moharil, Joaquin Vanschoren, Prabhant Singh, and Damian Tamburri. 2024. Towards efficient AutoML: a pipeline synthesis approach leveraging pre-trained transformers for multimodal data. *Machine Learning* 113 (July 2024), 7011–7053. https://doi.org/10.1007/s10994-024-06568-1

[33] Rishi K. Narang. 2013. *Inside the black box: a simple guide to quantitative and high-frequency trading* (second edition ed.). John Wiley & Sons, Inc, Hoboken, New Jersey.

[34] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*. ICLR, Online. https://doi.org/10.48550/arXiv.2211.14730

[35] Kenniy Olorunnimbe and Herna Viktor. 2024. Ensemble of temporal Transformers for financial time series. *Journal of Intelligent Information Systems* 62 (March 2024), 1087–1111. https://doi.org/10.1007/s10844-024-00851-2

[36] Cédric Poutré, Didier Chételat, and Manuel Morales. 2024. Deep unsupervised anomaly detection in high-frequency markets. *The Journal of Finance and Data Science* 10 (Dec. 2024), 100129. https://doi.org/10.1016/j.jfds.2024.100129

[37] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[38] Eduardo Ramos-Pérez, Pablo J. Alonso-González, and José Javier Núñez-Velázquez. 2021. Multi-Transformer: A New Neural Network-Based Architecture for Forecasting S&P Volatility. *Mathematics* 9, 15 (July 2021), 1794. https://doi.org/10.3390/math9151794

[39] Shosuke Sakagawa and Naoki Mori. 2022. Neural Ranking Strategy for Portfolio Construction Using Transformers. In *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*. IEEE, Phuket, Thailand, 95–100. https://doi.org/10.1109/IIAI-AAI-Winter58034.2022.00029

[40] Attilio Sbrana and Paulo André Lima De Castro. 2023. N-BEATS Perceiver: A Novel Approach for Robust Cryptocurrency Portfolio Forecasting. *Computational Economics* 64 (Sept. 2023), 1047–1081. https://doi.org/10.1007/s10614-023-10470-8

[41] Hugo Gobato Souto and Amir Moradi. 2024. Can transformers transform financial forecasting? *China Finance Review International* ahead-of-print (June 2024). https://doi.org/10.1108/CFRI-01-2024-0032

[42] Jifeng Sun, Wentao Fu, Jianwu Lin, Yong Jiang, and Shu-Tao Xia. 2022. Deep Portfolio Optimization Modeling based on Conv-Transformers with Graph Attention Mechanism. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Padua, Italy, 01–08. https://doi.org/10.1109/IJCNN55064.2022.9892317

[43] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. http://arxiv.org/abs/1409.3215 arXiv:1409.3215 [cs].

[44] Anh Tong, Thanh Nguyen-Tang, Dongeun Lee, Toan M Tran, and Jaesik Choi. 2023. SigFormer: Signature Transformers for Deep Hedging. In *4th ACM International Conference on AI in Finance*. ACM, Brooklyn NY USA, 124–132. https://doi.org/10.1145/3604237.3626841

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971 arXiv:2302.13971 [cs].

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, CA, USA, 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[48] Kieran Wood, Sven Giegerich, Stephen Roberts, and Stefan Zohren. 2022. Trading with the Momentum Transformer: An Intelligent and Interpretable Architecture. http://arxiv.org/abs/2112.08534 arXiv:2112.08534 [cs, q-fin, stat].

[49] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., Online, 22419–22430. https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf

[50] Bing Yang, Ting Liang, Jian Xiong, and Chong Zhong. 2023. Deep reinforcement learning based on transformer and U-Net framework for stock trading. *Knowledge-Based Systems* 262 (Feb. 2023), 110211. https://doi.org/10.1016/j.knosys.2022.110211

[51] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 9 (June 2023), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317

[52] Haohan Zhang, Fengrui Hua, Chengjin Xu, Hao Kong, Ruiting Zuo, and Jian Guo. 2024. Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements? http://arxiv.org/abs/2306.14222 arXiv:2306.14222 [cs, q-fin].

[53] Zhaofeng Zhang, Banghao Chen, Shengxin Zhu, and Nicolas Langrené. 2024. From attention to profit: quantitative trading strategy based on transformer. http://arxiv.org/abs/2404.00424 arXiv:2404.00424 [cs, q-fin].

[54] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. AAAI, Online, 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

[55] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Online, 27268–27286. https://proceedings.mlr.press/v162/zhou22g.html