

Original Article

Cross-Modal Economic Reasoning for Integrating Financial Text, Numerical Data, and Market Signals

Rudi Simanjuntak¹, Hendra Gunawan²

^{1,2} Syarif Hidayatullah State Islamic University Jakarta, Indonesia.

Received Date: 10 August 2026

Revised Date: 15 August 2026

Accepted Date: 20 August 2026

Abstract: Financial decision-making relies more and more on heterogeneous information, dispersed in financial reports, news articles, numerical indicators or time series such as price movements of stocks (stock prices), trading volumes and even planetary signals. Traditional analytic systems typically analyze these modalities separately, thus are incapable of discovering the connection between qualitative stories in economics and quantitative market activity. We propose a cross-modal economic reasoning network integrating financial text, numbers and market signals within the same analytical framework. We construct a model combining domain-aware language representations, temporal numerical encoders as well as market-signal extraction and cross-modal attention mechanisms which aims to capture relations among textual events, financial metrics with the dynamics of markets on them. Instead of considering different modalities as separate predictive variables, the framework focuses on temporal alignment, semantic correspondence between representations and causal modeling with evidence-weighted fusion. Recent work has shown that including both news and time-series information can enhance multimodal financial models, yet ongoing issues remain such as (1) temporal misalignment between modalities from the same data record; (2) modality imbalance in historical data collection for non-textual fields (e.g. sentiment); noisy/null modal inputs across records due to original channel sources/text parser errors; limited generalizability of model learned just using one market type genre to another unrelated domain/market type source only on flavor impetus eXchange application – we build bridges but need better ones! As a result, we propose that the framework incorporates modality-specific processing first and then uses adaptive cross-modal reasoning to enable the model to identify supporting, disagreeing and contextual signals. It provides a conceptual basis for more interpretable financial forecasting, economic event interpretation, risk mitigation and multi-reasoning demand decision-support systems over heterogeneous types of finance information.

Keywords: Cross-Modal Reasoning, Financial Text and Numerical Data, Market Signal Extraction, Multimodal Learning Methods, Economic Forecasting, Financial Analytics, Temporal Alignment of Market Signals, Attention Score Computation, Risk Assessment

I. INTRODUCTION

Financial markets are speed and high-dimensional information environments where economic decisions arise from the interaction of numerical measurements, textual data inputs. Feedback loops between investors' beliefs and market physics (regime changes) lead to a on-going shuffling(or esoteric competition in qualitative terms) They come in the form of financial statements, which offer structured quantitative evidence regarding revenues, profitability and different aspects (leverage & liquidity) or operational performance on one hand while contextual explanations for such numerical observations are typically found from news articles/analyst reports/regulatory announcements/(earnings communications/corporate disclosures). Conversely, the market price and trading volume (volume=size) at each time period can be interpreted as a reflection of how fast many traders react to whatever news is publicly available on the asset. These types of information are complementary but they do not typically co-exist in the same domains, at a similar refresh rate or be used for analysis purposes with an identical process. This has led to a key research problem of creating artificial intelligence systems that can reason across financial text, numerical data and market signals at the same time.

Historically, these types of information have been segregated by traditional financial analytics. While most quantitative models rely on numerical information – historical price and accounting data, macroeconomic indicators or technical variables – the natural language processing systems tend to centre around sentiment (i.e. removing variability introduced by linguistic opinion), event extraction, topic classification or semantic analysis of financial docs & news articles best suited for an English reader base. This separation can yield partial representations of economic conditions. For instance, the textual sentiment of an earnings announcement cannot account for its goodness in isolation. Whether it matters from an economic perspective could depend on revenue growth, earnings expectations, evaluation multiples and prior guidance or market gyrations as well as the price change in response to publication. In the same way, a strong price move could be statistically significant in a numerical model while its explanation has been restricted to an announcement regarding regulatory changes or corporate events. This means that cross-modal reasoning is not simply one of fusing datasets but



assessing the way in which information expressed within one modality affects how we interpret and predict from data expressed with another.

Recent developments in multimodal AI provide an excited basis to act on this challenge. There are two major aspects of financial multimodal research, deciding how to combine textual news with market time series or structured information in general, e.g., looking at tables and charts as common data sources. An extensive review conducted at the end of September 2023 provide a more detailed discussion about challenges in multimodal financial forecasting, classifying them as temporal misalignment; modality imbalance; missing or noisy information (such traditional textual data); and poor cross-market generalization. Evidence has also been advanced that language model research in finance should focus on models capable of working with tables, time-series data, charts and textual financial knowledge as opposed to pure text based inputs. These developments suggest an evolution away from single-modality prediction and toward the complicated web of financial intellect.

But the main hurdle goes beyond simply fusing multimodal data. Just natively concatenating also text and figure features will not guarantee economic reasoning. Financial data has temporal, relational and context dependencies. A news event can immediately affect a stock price for as long or even shorter than the time it takes to adjust upward or downward quarterly expectations; and, with an even longer lags over months, change investors perceptions of future risk. In the same vein, different market reactions can arise from repeated release of exactly the same sentence depending on valuation, macroeconomic conditions leading up to it as well broader sector sentiment or changing general investor expectations. Thus, a well-founded cross-modal system has to find the temporal alignment between sources of information and learn when one modality should enhance (or suppress) its counterpart or reinterpret this other modality.

An example case for this direction is a recent paper studying interleaved financial text and time-series sequences [11], which use modality-specific experts and cross-modal alignment to maintain the unique properties of each information type while performing joint causal reasoning. Other recent work includes hierarchical processing on language, where bidirectional interactions between text and price sequences have shown promise using co-attention to pair financial news with time-series of prices [5]. These results indicate that one could use cross-modal attention as a more sophisticated way to concatenate the static features.

This research will investigate cross-modal economic reasoning which is the systematic interpretation of relationships between financial textual references, numerical economical data and market observables. The framework frames each modality as an evidence source with specific advantages and limitations. Financial text adds semantic and contextual knowledge; numerical data gives measurable elements of economic & financial architecture, whereas market signals capture discovered behavior in the markets as well collective expectations. The interactions between these sources are then passed through a cross-modal reasoning layer that predicts agreement, contradiction and temporal causality as well as signal relevance within the utterances. It matters because financial decisions often hinge on competing signals rather than uniformly aligned data.

Temporal alignment is another important factor to keep in mind. Financial text may be published sporadically, accounting data may need to have a quarterly frequency, macroeconomic indicators can appear at monthly frequencies and market data is recorded on minute-level or even higher levels. This led to recent multimodal datasets emphasizing temporal alignment across financial news, structured financial tables and technical charts together with price series. Lack of temporal alignment leads to spurious predictive performance, because methods will often learn information correlating with market outcomes that could not have been available at the time of decision-making. Consequently, an effective cross-modal architecture needs to maintain the accessibility of information and form temporal alignments between signals (i.e. outputs) alongside modalities such as time-stamped data).

Another motivation for the research is interpretability. Financial forecasting models no longer need only to be able to scribble down big figures. Investors, analysts, financial institutions and regulators want to be shown what evidence led to a conclusion interpretation from which data has the textual information demonstrated in numbers or market. Such interpretability can be facilitated with cross-modal reasoning, which is also able to point out the major financial statements and their numerical changes responsible for driving market movements as well as interactions among them. This is especially important for risk-sensitive applications where unexplained predictions can be hard to validate or put into production.

II. CROSS-MODAL ECONOMIC REASONING ARCHITECTURE

To integrate the financial text, numerical data and market signals one needs an architecture that preserves unique characteristics within each modality while building relevant relationships between them. Financial information is a heterogeneous collection: textual sources provide semantic and contextual data, numerical datasets describe measurable

economic states while market signals reflect dynamic responses by investors and other institutions. A traditional machine-learning architecture, which merges these inputs into a single feature vector, may miss the temporal and semantic dependencies that govern their economic relevance. Consequently, the architecture of cross-modal economic reasoning should structure information sources to be represented independently and only aligned, compared or jointly interpreted at a later stage.

The proposed architecture has three key stages: (i) modality-specific representation; (ii) temporal and semantic alignment followed by cross-modal reasoning via adaptive fusion. In the first stage, financial documents and market narratives are converted into contextual embedding's while numerical financial variables encoded in accordance to their temporal and economic properties. Processed through task-specific temporal representations, market signals such as returns, volatility, trading volume momentum liquidity and order-flow indicators. We link these representations during the second stage by publication time, observation frequency and reporting period lengths along with event windows. The last stage uses the cross-modal attention in a sense to see how much information of one modality is modifying or can modify the interpretation of another.

As such, this architecture can differentiate between positive and negative evidence. For instance, enhanced textual sentiment alongside worsening profitability and poorer market prices should not necessarily be seen as a bullish condition. Yes, the model need not know all modalities are conflicting and that it has to learn the contextual relation amongst them. Additionally, positive earnings surprises plus negative market reactions may signal that prior expectations have been baked into the stock. Therefore, cross-modal economic reasoning builds a network of proof from the not uniform assortment of monetary data rather than treating single perceptions as lacking prescient capacity.

The architecture is also aligned with explainability principles, by keeping modality-level contributions intact in steps. The advantage of this is that one final prediction can be related to textual events, numerical fluctuations and the market. Hence, their proposed method offers an even conceptual roadmap toward context-sensitive temporal and economic AI systems.

A. Modality-Specific Financial Representation

Modality-specific representation learning is the first part. Because of the difference between grammar and rules governing structured documents or financial text (and numerical data, market signal), it necessitates unique encoders. Financial text may consist of earnings announcements, annual reports, regulatory filings, analyst commentary statements made by central banks and financial news. A domain adapted language encoder can convert such documents into deeper regional contextualized vectors containing sentiment, financial entities, events with relations and also economic semantic meaning. However, the representation should separate statements of historical performance from those relating to current conditions and expectations as of a prospectus date.

B. Temporal and Semantical interoperability of heterogeneous data

To this end, temporal and semantic alignment is an important part of cross-modal economic reasoning because financial information exists at non-overlapping time frames as well as frequencies. For instance, a company can release an earnings signal at some specific time stamp but market prices could be sampled every minute and accounting indicators might describe an entire quarter. If these observations are misaligned, a model can 'leak' information from the future associated with past market conditions leading to unrealistic predictive performance.

C. Cross-Modal Attention and Adaptive Evidence Fusion

Cross-modal attention and adaptive evidence fusion: the final part of architecture Once modality-specific representations have been obtained and aligned, the model should be able to decide how much weight each information source should contribute towards the final economic interpretation. Static feature concatenation makes the assumption that each and every modality contributes equally, which hardly is a valid one in financial settings. For example, a financial news article may have high importance just after the announcement but will be much less significant several weeks later.

This limitation of simple self-attention is mitigated in cross-modal attention, where representations from one modality can attend to information contained within another. For example, a textual representation can attend to respective changes in financial variables whereas market representations are able to map into news events that potentially explain the corresponding observed price movements. As an objective measure of whether a narrative is backed by financial performance, numerical indicators can even influence how textual information is interpreted.

III. MULTIMODAL DATA HANDLING AND ECONOMIC SIGNAL INTEGRATION

The information that financial markets generate comes from several channels, whose structures, frequencies and reliability differ greatly. Business conditions are summarized qualitatively by financial reports and news articles, but

numerical financial data quantifies organizational performance or macroeconomic strength. Market signals, represented by price movements and volatility in addition to volume, liquidity or additional order-flow indicators is the ongoing response of market participants having access to information. It is insufficient to simply plug heterogeneous covariates into a single prediction model across these disparate datasets. One needs to train the system on how textual events relate with each other and possibility of numerical changes (and vice versa), showing subsequent behavior in market; both cause and effect.

This solves this problem by creating a multi-modal processing architecture that creates an entire pipeline of raw information ingestion to unifying economic reasoning. It starts with the acquisition of financial documents, amenable structured datasets from management accounts and market time-series observations as well as relevant context indicators. Subsequently, we process each modality with a proper representation method. Natural language processing approaches can extract financial entities, events from text sentiment, uncertainty and forward-looking statements. Computational techniques for financial variables normalize factors, trend findings technological changes and relationships among economic measures. Time-series models can model market dynamics and also be used to detect changes in volatility, momentum, liquidity and trading behavior.

The main purpose is to ensure that samples need not be directly compared, but rather only embedded representations must keep modality-specific information and at the same time should make cross-modal relations measurable. The reason its temporal alignment is so important, in particular because each information will be released and time. Quarterly results from a company are released at specific times of the year, which can leave its stock price and trading volume vulnerable to rapid change. These observations give the model a chance to look at, whether there is inconsistency between announced information and market responses.

The resulting framework accommodates multiple financial intelligence applications, such as market forecasting and event interpretation; risk assessment and anomaly detection in stocks; portfolio analysis for investment under complex nonlinear dynamic systems (including high-dimensional multivariate time-series data); economic decision support. Using semantic evidence, we propose to integrate this information with quantitative and market evidence in order to create better context-aware explainable financial reasoning.

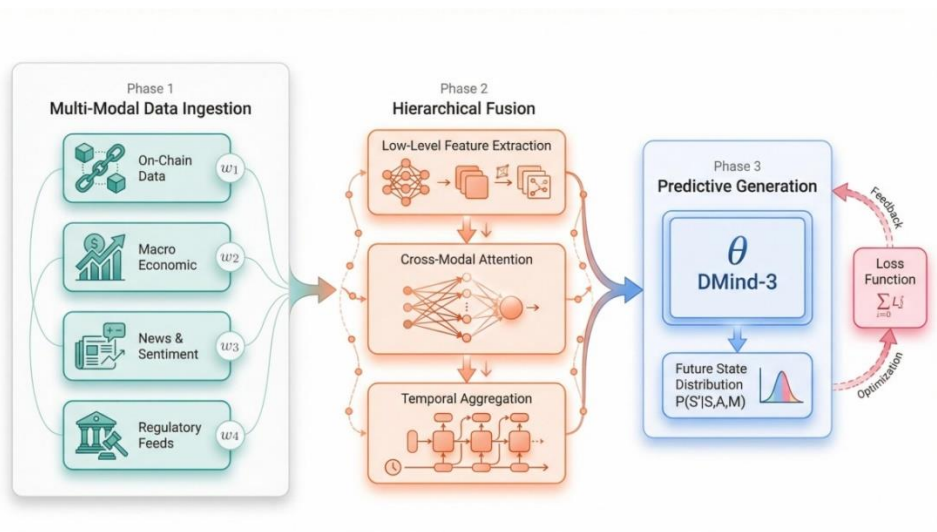


Figure 1: Multimodal Financial Information Processing and Economic Signal Integration Framework

The initial phase of multimodal financial intelligence is the conversion of inconsistent and non-standardized layers or types of financial data into machine-digestible formats. For instance, the following inputs can be categorized under financial text; annual reports, earnings announcements (filings), regulatory disclosures or notices of violation against principal agent(s) acting on behalf a firm, analyst report(e.g. Home Depot stock rating news), central-bank and macroeconomic communication structures in speech formats as well corporate statements which are composed primarily by internal stakeholders responsible for image-repair strategies when reputational crisis breaks out - nor do they mediate between all these competing narratives? The papers involve the financial performance, management expectations information, administration recommendations including magnitude of connections and such market events related to it. Current language-processing models can convert such information into contextual representations while recognizing sentiment, entities, financial events (including qualitative and quantitative predictive indicators), relationships between them, as well as forward-looking statements.

Financial information, on a numerical basis, is processed on another level as it gains its meaning in terms of scale (big or small), timeframes (short- and long-term) and relations between more variables. Some of the important variables could be revenue, earnings and cash flow along with debt-loads, profitability metrics (such as return on equity), inflation rates policy interest rate environment exchange rates GDP growth differential valuation ratios and financial risk measures like Altman Z-score. Rather than viewing these variables as numeric inputs that are independent of each other, the framework proposed represents trends or changes in those relationships over time. So what a 10 percent increase in revenue means could be significantly different if operating costs jumped up by, say, to 20 over the same time period.

Once the independent representations are obtained, we can then build relationships between modalities in a second stage. Why cross modal alignment is necessary: Financial information does not come at a common frequency. News can go out at any time, accounting reports come quarterly (and all the rest), macroeconomics data is monthly and market-sensitive information second-, minute- or daily level intervals. This framework links information by publication time, reporting period, event type and window for unconventional market responses. Now think of an earnings announcement with revenue strongly beating expectations— The text representation may show sentiment as positive, the numerical presentation can identify revenue growth and better margins while market signals reflect a price rise immediately followed by increased trading volume. The system considers these observations corroborating evidence when they agree. On the other hand when good textual information is followed by prices declining and rising volatility, a proper model should see an inconsistency instead of assuming bullishthe announcement.

This attention method allows each modality to selectively pay attention to information coming from all modalities. Text features may attend to numbers, numerical representations can be categorized against market reactions and marks can cue which financial events might explain their movements. Adaptive fusion mechanisms can then particularly evaluate the individual contribution of each modality.

IV. CROSS-MODAL ECONOMIC REASONING AND FINANCIAL DECISION INTELLIGENCE

This supports the creation of financial intelligence systems that combine text, numerical data and market signals using heterogeneous evidence reasoning. Yet, multimodal integration is not sufficient to produce economically meaningful results. A well-structured system will often combine many pieces of information but still fail to model when the data is relevant, and how that relevance changes over time in terms of contradicts or dependencies -- it just organizes multiple sources together. This limitation of approaching the evidence modality-specifically is addressed by cross-modal economic reasoning, where multimodal representations are converted into a interpretable decision process that aggregates multiple types of financial evidence.

Most economic information contains relationships that are not identifiable from a single modality. A positive earnings statement by a company can look good on paper, but if its revenue growth is below what analysts expected to see or the price of share drops post-publication it be meaningfully different. Likewise, an increase in trading volume could mean that investors are confident, confused, institutional repositioning or market reaction to a news item/incident. Contextual evidence from multiple sources is needed to figure out which interpretation works.

A. Multi-Modal Interpretations of Economic Events

Economics events are critical moments where text, states and behavior overlap. This could cover anything from earnings announcements, mergers and acquisitions, interest-rate decisions to inflation releases/regulatory changes/ credit-rating adjustments/geopolitical developments or corporate guidance revisions. Such events should be interpreted by a cross-modal system that assesses the evidence across all available information channels and not only based on text sentiment or historic price patterns.

The stated method identifies pertinent occurrences from monetary reviews by financial named-entity recognition, event extraction, semantic classification and contextual language representations. Then the observed event is related to several numerical descriptors and observations, as in a deep market for selling. An earnings announcement can be related to estimated sales, EPS numbers, changes in revenue or operating margins over time, analyst stock target estimates and the movement of associated salaried stocks information capturing price performance metrics such as average 1 month cumulative return / volatility relative remapping into a full-day volume observation period ($t+2$).

B. Temporal Reasoning and Market-Response Analysis

A key property that financial information has is it matters for different time horizons. Market reaction to an announcement can take place in seconds or minutes; the impact of that announcement on earnings expectations, investment behavior, and valuation may last for months. Thus a cross-modal model needs to learn not only what information co-relates with the market movement, but also when it can be accessed and how influential is that information over time.

The architecture outlined creates temporal windows centered on events of financial materiality. Such windows may encompass pre-event market behavior, immediate post-event response and longer-term adjustment. This segmentation allows the system to compare market conditions before and after information is made public.

C. Contradiction Detection, Uncertainty and Evidence Reliability

Financial facts often include mixed signals. Management can prescribe optimism while profits are pulling back, a company might miss on earnings but offer better than expected guidance and its stock price falls anyway, or key economic indicators may show signs of improvement at the same time that market volatility escalates. This kind of scenario exemplifies why some financial AI implementations should never take for granted that all modalities provide homogeneous evidence.

Cross-modal reasoning: representations across modalities can explicitly indicate a contradiction. You can compare textual sentiment with numerical performance, and market responses could be compared to both. Relatedly, one could create a conceptual definition for the contradiction score as some function of 1) semantic disagreement in claim content, 2) numerical divergence between expectation and market-response magnitude. Uncertainty estimation is equally important. Financial text might be written in an ambiguous manner, numerical datasets may have missing observations and market signals might become distorted by short-lived volatility. Instead of pulling together a very high confidence prediction from weak evidence, the system should be likely to estimate its confidence based on not only quality in data for one modality but also agreement or disagreement between modally cross-modal features.

The framework, enables to classify evidence into supporting, neutral contradictory and uncertainty. This happens when multiple modalities are suggesting the same interpretations. Contradictory evidence is when the modalities predict a different result. Uncertain evidence comes into being when we have incomplete or unreliable information. These capabilities enhance the robustness of financial decisions-system. This does not work in isolation — rather than only outputting a prediction, the model can specify whether that prediction is endorsed or conditional on sources of conflicting evidence. This is useful primarily in investment research, managing financial risk or asking questions at an institutional level.

D. Explainable Financial Decision Intelligence

The last step for cross-modal economic reasoning is converting model outputs into readable financial diamonds. In contrast to traditional predictive models, which may spit out a prediction without transparently elaborating how small pieces of information might have been informative for the conclusions. This constraint towards validation can breed distrust in financial environments, particularly to analysts who independently happen upon these disclosures. It has an explainable cross-modal architecture that maintains modality-specific evidence during the reasoning process. A prediction can therefore be mapped to relevant financial statements, numeric variations in the time period of interest for a stock or market indicator associated with it and any detected economic event that triggers such an analytic (figure 3). For instance, a model predicting higher market risk might flag volatility moving up while liquidity deteriorates alongside negative financial-news sentiment and popular vagueness in the news (uncertainty-related language hitting highs).

V. EVALUATION FRAMEWORK AND COMPARATIVE ANALYSIS

In order to evaluate a cross-modal economic reasoning framework, it is not sufficient to only measure the conventional prediction accuracy. The suggested system includes financial text, numerical features and market signals therefore in order to evaluate its performance it has to be seen from different aspects: predictive effectiveness cross-modal alignment temporal coherence robustness interpretability computational efficiency. For example, it is possible that a model predicts well but does not recover the causal link between an economic event and its market effects. Similarly, an explanation may be very interpretable but the model that produces it can perform badly when market conditions change. Hence it is expedient that the most comprehensive evaluation framework incorporates both quantitative and qualitative dimensions.

You will first start by splitting the financial information into datasets (by modality) and building synchronized multimodal observations. The second subtask is to evaluate financial text, which can be accomplished with event extraction, sentiment classification and semantic relevance (scoring their similarity) as well as financial entity recognition. Numerical facts can be summarized in a similar representative fashion through trend representation, financial-ratio relations and forecasting variables. Return, volatility, trading volume/ liquidity (traded proportion) and momentum are all wsp_{500} or time-series based market signals. Ablation analysis should be a significant part of evaluation. Researchers are able to find whether a reduction in one of the modalities, such as financial text or numerical information or market signals leads to meaningful incremental value. We can then use these comparisons between unimodal, bimodal and fully multimodal configurations to ascertain whether integrating cross-modal inputs really enhances economic reasoning.

Table 1: Compact Evaluation Framework for Cross-Modal Economic Reasoning

Evaluation Dimension	Key Measures	Purpose
Predictive Performance	Accuracy, F1-score, MAE, RMSE	Measures forecasting and classification quality
Cross-Modal Alignment	Alignment score, retrieval accuracy	Tests relationships between modalities
Temporal Robustness	Rolling-window performance	Evaluates stability across market periods
Ablation Performance	Modality removal comparison	Measures contribution of each modality
Contradiction Detection	Precision, Recall, F1	Tests identification of conflicting evidence
Explain ability	Attribution consistency, expert evaluation	Assesses quality of model explanations
Robustness	Noise sensitivity, missing-data performance	Tests resilience to imperfect inputs
Efficiency	Latency, memory usage, computational cost	Evaluates practical deployment suitability

The table shows that the proposed framework should be evaluated more as a complete financial intelligence system rather than as merely forecasting. Predictive performance measures whether the system can produce useful outputs, while alignment & ablation establish if multimodal integration contributes any useful information. It assesses the temporal robustness of a system by checking if market conditions change, and it also tests contradiction detection where the ability to reason is challenged when two available sources are contradictory. Ability and robustness clarify practical needs for reliable use, while computational efficiency defines if the architecture can run inside real exchange analytical domains or not.

VI. CROSS-MODAL ECONOMIC REASONING IN FINANCIAL INTELLIGENCE

Cross-modal economic reasoning can leverage the knowledge gained cross-modally to create variably intelligent systems which assemble heterogeneous information into an interpretable form, and thus may catalyze transformations in financial intelligence. Many financial applications are built with dedicated models for specific tasks like sentiment detection, pricing prediction or credit scoring and statement analysis. These systems often work well for the specific domain they operate in but may fail to take into account dependencies between information sources. A cross-modal framework offers the abstraction which puts together financial narratives, structured economic factors and observed market behavior in an integrated reasoning engine.

This integration has become especially relevant due to the growing availability of financial data. The average organization today produces and ingest vast amounts of annual reports, regulatory filings, earnings announcements, financial news articles/economic statistics/market prices/trading activity/refinancing textual data along with alternative finance indicators. Now the challenge is not so much in acquiring information but figuring out what information is important, how different sources relate to each other and how that relationship evolves over time. This problem of integrating semantic, numerical and temporal representations is tackled by the cross-modal economic reasoning.

The applications can be as varied from market forecasting, investment research, financial risk management, portfolio decision support to economic event analysis and institutional intelligence. A forecasting system may therefore use a combination of sentiment from the financial press and highlights among company fundamentals, as well recent price behavior. A risk-management framework can detect detrimental corporate disclosures alongside monitoring the market's volatility and liquidity environment. Investment-analysis model compares management assertions with actual market performance and outcome expectations

A. Cross-Modal Financial Market Forecasting

Cross-modal economic reasoning is extremely useful in finance, especially when it relates to financial market forecasting. Traditional forecasting is based on the time series of previous prices, technical indicators and statistical relationships accordingly to multiple examples. While these variables have valuable information, the price action is often driven by external events and investor interpretation of those other factors. Market movements can be influenced by information provided in financial news, corporate announcements, and key economic policy actions or statements that lag the release of pricing data which cannot all be explained with just historical price patterns.

A cross-modality forecasting framework which we term as the MTF combines multi-modal market time-series with textual and numerical financial features to address this limitation. The textual encoders recognize sentiment, events, uncertainty and forward-looking statements while numerical sensors process company fundamentals along with macroeconomic factors. A market encoder captures these price trends, volatility, volume, liquidity and momentum all at once. Its role is to ascertain the interaction between these information sources prior to forecasting using cross-modal attention.

B. Investments and Portfolio Decision Intelligence

Investment research for example can become more compelling using cross-modal economic reasoning by allowing analysts to evaluate companies and financial assets based upon qualitative as well quantitative evidence. Common aspects that determine the decisions around investments include financial statements, management commentary, industry news analyst expectations valuation metrics market performance and macroeconomic conditions. It can be difficult to have a clear investment thesis when processing these sources separately.

By relating predictions in a claim to quantitative facts and market actions through natural language, it can build a unified representation of an investment. Management may talk about driving operational improvements, but operating margins and cash-flow performance provide the proof. Then market signals can tell you whether investors have responded or expectations are still on the sidelines.

C. Financial Risk Management and Economic Events Discussion

In financial risk management, it is important to continuously monitor the information that may suggest a transition in credit, market, liquidity system or operational risks. Traditional risk systems usually focus on quantitative measures such as volatility, leverage, credit spreads or other liquidity metrics and historical losses. Nevertheless, crucial risk information often first emerges from textual sources such as regulatory announcements, corporate disclosures, litigation reports and management statements and financial news. Cross-modal economic reasoning bridges the gap between these textual indicators and numerical and marker evidence. For example, an announcement by a company that it expects cash flows will be negatively impacted in the future while its liquidity is dwindling and volatility surging. Any single one of these signals might not hint at the same degree of deterioration in financial conditions as this combined evidence.

The system is also capable of detecting economic events itself by merging time-series economic text with a dataset containing high-frequency numerical indicators and market reactions. Gross domestic product, durable goods orders and consumer spending could all be measured against interest rates, exchange rates (including FX store-of-value assets) equity prices commodity prices volatility measures as central-bank announcements inflation releases employment reports geopolitical developments other regulatory changes etc.

VII. CHALLENGES, LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Cross-modal economic reasoning is a good step towards achieving human-like financial intelligence, but it brings with its own set of methodological, technical and economical challenges. The heterogeneity of financial information goes beyond the form; it also involves the source, frequency, granularity and interpretation. Text has inherent text ambiguity & imbalance → Numerical data can have missing / inconsistent observation → Signal (Market) is not clean, precise – affected by noise/ liquidity conditions/investor behaviours/unforeseen events Therefore, these sources need mechanisms that can identify valuable economic evidence and separate it from irrelevant or misleading information.

Perhaps the biggest challenge comes with getting things to play nicely in time. News about finance is milliseconds old compared to the market reaction; accounting information covers past periods and macroeconomic indicators have a regular publication schedule. Improper alignment of a time component can result in leakage. A second challenge is that of modality imbalance; one information source may dominate the learned representation, leaving other modalities contributing little. This means making dynamic judgments on how reliable and relevant each source of information is.

Generalization is another important limitation. That contrasts financial ties can be across companies, industries, nations and market environments. A model that works well in times of normal market conditions fails at resilience during a financial crisis or unusual periods where volatility is extreme. In addition, the inherent differences in financial reporting practices and market structure such as cultural specialties & understanding cost of capital between two markets or regions can further diminish cross-market transferability. Interpretability and computational cost still play a major role. Financial institutions (for example) often work on low-latency signals so large multimodal models will require a lot of computational resources. Also, banks may need to transfer data back and forth between various servers but that can be reduced if the pipelines are lightweight in terms of multi-modality. At the same time, complex models must give clear enough explanations for analysts and decision-makers.

Table 2: Challenges and Future Research Directions for Cross-Modal Economic Reasoning

Challenge	Current Limitation	Future Research Direction
Temporal Alignment	Different data frequencies	Point-in-time and event-aware alignment
Modality Imbalance	Dominant information sources	Adaptive modality weighting
Data Quality	Missing, noisy, biased information	Reliability-aware preprocessing
Market Regime Changes	Reduced generalization	Regime-adaptive architectures
Interpretability	Complex multimodal decisions	Evidence attribution and causal explanations
Computational Cost	High training and inference requirements	Efficient multimodal models
Cross-Market Transfer	Geographic and structural differences	Domain adaptation and multilingual models
Data Privacy	Restricted financial information	Privacy-preserving learning

A. Data Quality, Temporal Alignment and modality balance

A cross-modal financial reasoning system is significantly only as good as the input data combination quality and time synchronization. Financial text can be ambiguous, copy & pasted reporting style or even contradictory interpretation. Numeric datasets are more likely to suffer from missing values, revisions of data used for accounting purposes, different units or delays in reporting. Market data: will generally include temporary anomalies, extreme observations and microstructure noise. This can for example be a problem when these problems manifest simultaneously in the data and thus the model learns relationships that are not actually true economic behavior.

Temporal Alignment need is the most difficult task here due to 3 major modalities are working on different frequencies. Financial statements could be at the quarterly level, economic indicators are monthly and observations on market movements both for intraday. Thus, future research should establish event-aware temporal alignment mechanisms that maintain the precise availability of information at each prediction timestamp. Modality imbalance presents another problem. The amount of numerical market data might overshadow much smaller textual datasets leading the model to become heavily reliant on numeric signals instead. Adaptive weighting and modality-specific regularization could keep contributions to a minimum.

B. Generalization, Explain ability, and Robustness

This is true in the context of a financial model that needs to work when economic conditions change. As an example, relationships that you learned during a bull market may not hold up in a financial crisis and one sector might show patterns shall we say that do not necessarily apply to another. Research should therefore focus on developing regime-aware models which have the ability to identify shifts in volatility, liquidity conditions and investor behaviours. Explain ability is equally important. Cross-modal models can be rapidly become quite intricate due to the fact that several attention layers and also fusion systems contribute in making a particular prediction. There should be evidence-level explanations for future systems attesting which textual statements, numerical indicators and market signals led from an input file to a decision. The robustness should also be studied under imperfect information, e.g., the missing and contradictory observations. Realistic systems ought to be fault tolerant when one of the modalities is unavailable for a time period. Learning to mitigate modality dropout, uncertainty estimation and reliability-aware fusion are some areas of research that could enhance resilience.

C. Efficient models, privacy and future directions

The real-world deployment of the multimodal financial models can be obstructed by their high computational requirements. Large language models and complex time-series architectures can be computationally demanding, especially when financial data needs to be extracted iteratively. In the future, research can focus on parameter-efficient learning (based others [80], heteroscedastic adaptation methods like [66]), model compression techniques such as distillation for some of latest natural language models (5), and more efficient attention modules to increase relevance in financial time-series analyses([38]). Privacy is yet another issue because financial institutions may have proprietary data sets that cannot be freely shared. Collaborations can be facilitated by privacy-preserving learning techniques which help to minimize the need for direct exposure of sensitive finance data.

Further research can even focus on causal cross-modal reasoning. Existing systems can pick out patterns between news, numbers and market movements without determining whether one actually led to the other. Making use of

multimodal learning and causal inference together could thus give much deeper economic motivation. Finally, future frameworks ought to include alternative data sources — such as multilingual financial information, ESG disclosures, supply-chain indicators social sentiment and real-time economic events. These expansions might carry cross-modal economics reasoning out of prediction analytics into financial intelligence that engages time series with real-time interpretation for adapting finances to unfolding economic environments.

VIII. AN ADVANCED METHODOLOGICAL FRAMEWORK FOR CROSS-MODAL ECONOMIC REASONING

If one is going to reason economically, it has first of all to be possible before reasoning cross-modally: different types or modalities of financial information must lend themselves with measure and representation that facilitates comparison against itself mode(s) in a common analytical environment. We introduce an economic reasoning based framework which generalizes conventional multimodal learning by adding a layer between data fusion and prediction. This difference is significant because the financial information does not simply exist across modalities; rather, it interacts hands-on. A statement in a financial report could affect how we interpret the same numerical indicator while a market response can be used as evidence of what investors believed about that interpretation.

The following sections describe the methodology principles at a high level, where an outline of our proposed methodology framework is represented through four interrelated components: (1) multimodal feature engineering, (2) cross-modal representation learning, (3) adaptive economic reasoning and finally ending with decision-oriented output generation. The first part is the preparation of financial texts, categorical variables and signals from market dynamics while preserving their uniqueness. The second part aligns these diverse representations to a common latent space. The third component disks attention among modalities, weights evidence from the source modality for tasks such as contradiction detection and temporal reasoning. The last step converts the encoded representation into predictions, risk measures, event classifications or decision-support-shaped outputs.

A. Multimodal Financial Feature Engineering

Stage 1 — Raw financial data is organized and then relationally represented in a way that can be mapped across modalities. The entities, events and sentiment/uncertainty of financial text can be extracted from the data to build models for use in natural language processing (NLP) tasks. For instance, through normalization, logarithmic transformations (for non-linear relationships), ratios, growth rates and rolling statistics using temporal differencing for numerical values. Market signals can be captured by returns, volatility, volume momentum liquidity and other types of time-series factors. The goal here is to keep the economic rationale behind each modality instead of considering all inputs as same feature. Contextual relationships should be maintained with textual representations, and numerical ones should maintain both magnitude as well direction. Any representations that markets withhold should preserve temporal dependencies.

Feature engineering can also add contextual variables like industry classification, company size, macroeconomic conditions and market environment. This gives more detail on how to understand why identical signals might produce different outcomes in some circumstances. A proper preprocessing phase should also catch lacking observations, duplicated values, outliers and inconsistent timestamps. The processing of information related to point-in-time, is also critical as it should contain the data only up until a certain timestamp of when the prediction is made publicly available.

B. Cross-Modal Representation Learning

Post-preprocessing, the framework produces modality-specific representations and sets relationships between them. An encoder for financial language can work to encode information that is textual in nature while numerical and market encoders will learn structured and time-series patterns. And these representations are encoded into compatible latent spaces. Interaction is enabled through cross-modal attention. For example, a textual representation can attend to relevant financial variables; market representations are able to pay attention similarly to news events happening contemporaneously. Such bidirectional interaction makes it possible for the model to confirm whether information source If a particular input is reinforcing or contradictory.

Take, for example, a text that states business conditions are improving it can be compared by real data on revenue growth at operating margin and market performance. If all sources suggest improvement the representation gets more support. Whereas simple averaging would eliminate these contradictions, the model is able to preserve them (for example if once the entire sentiment expressed by text in a room explanation is net positive and yet perform poorly on its own). The representation thus includes modality-specific as well as cross-modal information. This makes it possible for the latter reasoning layers to freely analyze financial situations from different angles.

C. Decision-Oriented Prediction and Evaluation

The last step of the methodological approach converts these two complementary and integrated economic representations into task-related outputs. The model can create the predictions — such as whether to buy or sell (market-

direction), if a stock is going up or down, any potential volatility factors over certain periods of time, financial risk scores based on specific criteria and so forth via event classification into multiple classes accompanied with an alarms that alerts users when fluctuations occur. Depending on whether the problem is regression or classification based, some possible metrics for forecasting are MAE, RMSE, directional accuracy (DA), precision recall and lift chart F1 score area under ROC curve. When conducting hands-on explorations of investing strategies, financially meaningful metrics such as cumulative return, maximum drawdown, Sharpe ratio and risk-adjusted performance can also be easily integrated into your financial applications.

Evaluation compares the proposed cross-modal framework with both unimodal and bimodal baselines. For instance, we can run a text-only model against the numerical-only model and its market-only counterpart as well as our text-plus-market variant on top of all three components. Ablation experiments can be used to determine the role of each component. We can determine whether cross-modal attention, adaptive weighting, temporal alignment or contradiction detection yields benefits by removing a given component from the mpt.

IX. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

To evaluate a cross-modal economic reasoning framework, an experimental design that captures the complexity of financial information and constantly changing nature of financial markets is required. As the proposed approach combines financial text, numeric economic data as well market signals for forecasting purposes we must not only assess whether this system improves prediction accuracy but also if indeed it successfully captures associations between heterogeneous information sources. This is why, we argue for an experimental design that will separate real cross-modal reasoning and what can be gained by merely adding to the volume of inputs.

The starting point of the experimental procedure ought to be a coherent multimodal financial dataset. For example, textual information can be corporate announcements, financial news and annual reports; regulatory disclosures such as Form 8-K filings; or analyst commentary. Numbers can be in the form of financial statements or accounting ratios and even macroeconomic indicators or economic statistics. Market information such as historical prices, returns, volatility, trading volume and measures of liquidity or other metrics derived from the market. These datasets need to be synchronized, i.e. aligned according to when the information was disclosed into public domain (e.g., it is likely that not all companies will unveil their earning announcement on exact same date).

A. Multimodal Dataset Construction and Experimental Setup

Constructing a multimodal corpus with good quality is of great importance for an experimental evaluation. This new dataset would feature three major types of information: financial text, structured numerical data and market signals. Your data can be gathered from company announcements, earnings release documents, regulatory filings during the end of a quarter/year period collected through financial newspapers/analysts and economic-policy communication. The key variables for the model can include revenue, earnings, profit margins and ratios (such as leverage or liquidity), valuation multiples/ratios across countries (including inflation, interest rates such as treasury yield curves; GDP indicators in each country. Market signals can take the form of prices, returns, volatility and trading volumes to spreads and momentum indicators.

Your observations should be timestamped, indicating when the data was available to you. It is required to avoid look-ahead bias. Do not use text published after a market movement to explain that earlier market movement. A dataset may be structured as a set of event-centered samples, where each particular important financial occasion is associated to prior numerical situations and future responses in the market. Outliers, data that is missing or having repeated entries in reports should all be taken into consideration with a systematic approach.

B. Baseline models and experimental design for comparison

A comparative baseline structure is needed to understand if cross-modal integration offers unique benefits. The first type comprises unimodal models. A financial-language model can handle text data, a numerical forecasting model can account for structured inputs, and a time-series model accounts for all those market signals. These models define the predictive capacity of single modalities. The second includes bimodal models (text-plus-numerical, text-plus-market and numerical-plus-market). We run these experiments to know if using two sources of information gives better results when combining.

C. Performance and Financial Evaluation Metrics

The evaluation performance should include both the classic machine-learning metric as well as concentration metrics related to finances. For classification problems like predicting market direction or identifying economic events, you can rely on accuracy/precision/recall/F1-score & area under ROC. For continuous forecasting tasks, prediction quality can be measured as the regression error (MAE, MSE or RMSE). In contrast, financial applications will require further validation since economic relevance does not automatically follow statistical accuracy. It can be the ratio of times that the model has

correctly predicted any market move direction (the "direction accuracy"). The addition of aggregated return, Sharpe ratio, maximum drawdown and volatility across the backtests might serve as enough evidence for practical effectiveness when predictions are employed in simulated investment strategies.

D. Ablation Studies, Robustness Test and Generalization

Although the proposed framework is remarkably simple, an important follow-up study would be to conduct ablation analysis rigorously to determine which specific part of the framework helps in its performance. The full architecture can be compared against variations where each of financial text, numerical data and market signals is removed independently. We can further remove (a) temporal alignment, (b) cross-modal attention (c) adaptive evidence weighting or (d) contradiction detection with the experiments. In robustness testing, we should test how the model performs when information is missing or sometimes noisy. As an example, a textual input can be partially elided, for numerical variables they could be missing observations or market signals denote a fractious volatility. It should fail gracefully instead of failing if one information source is missing in a reliable cross-modal system.

X. CONCLUSION

Cross-modal economic reasoning is an important line for building better generalizable financial intelligence systems. The paper considered a framework to combine three fundamentally different (but complementary) information sources, namely financial text, numerical economic data and market signals. Typically, these modalities are processed separately using the traditional financial analytics that often fall short in identifying how qualitative economic narratives relate to measurable financial conditions and actual market behavior. This limitation can be addressed by posing financial intelligence as a multimodal reasoning problem instead of merely breaking it down to prediction problems.

The final proposed solution model is formed from several domains flowing composed stages: 1) modality-specific representation, 2) temporal alignment and semantic alignment for training cross modal attention to reduce the complexity of watch list tables in RV, 3), a robust adaptive evidence fusion mechanism via contradiction detection over evidence generation/contrast points. Financial text has been adding some semantic information regarding events, sentiment, expectations and uncertainty as well as whether there is any downside to a corporation or economy. While numerical data gives tangible proof of financial outcomes and macroeconomic trends, market signals record the ever-evolving perceptions of investors and institutions alike. By combining these sources, the system can create a fuller picture of what is happening in the lower-level economy.

One main take-away from the proposed conceptual framework is that multimodal financial intelligence (FI) cannot be achieved just by concatenating features from different datasets. Some information is available at different frequencies and some data goes back to other times — this requires temporal alignment. Interpretation of models can lead to data leakage if not handled properly and will provide unrealistic predictions. Likewise, adaptive and evidence-weighted combination is a must since the relative significance of each modality varies by event incident, market regime variation from month to quarter year forecasting horizon. The paper also stated that it was essential to reason under conflicting evidence. Financial data does not always provide a reliable signal. Good management commentary may coincide with falling fundamentals, and good results at behind bad ones because the expectation of investors was already parked on prices. Thus, a strong cross-modal system should detect supporting evidence, contradictory evidence along with neutral and uncertain indicators — while not assuming that all the information sources reinforce each other.

This framework can be implemented for financial market predictions, investment research and portfolio decision support metrics, risk management applications. From the way economic events affect a multitude of intricacies All this leads into anomaly detection The universe in which these data are sourced Financial intelligence Cross-modal reasoning can establish connections between news events, financial fundamentals and price behavior in the realm of market forecasting since forecasts are usually based on time series data from various cross-sectional sources such as historical prices for stocks or commodities over a fixed timeframe. In investment research, it could help analysts compare qualitative corporate data to quantitative financial performance. It can detect early warning signals scattered across financial disclosures, numerical indicators and market activity in risk management.

Further, the evaluation framework we propose in this paper also gives a strong argument that cross-modal financial systems need to be analyzed and evaluated by multiple dimensions. Predictive performance can be measured using accuracy, precision, recall and F1-score; while financial measures such as directional accuracy, Sharpe ratio maximum drawdown cumulative return could evidence economic usefulness. Ablation experiments are especially crucial as they formally test if each modality and reasoning mechanism conveys non-negligible additional information.

Cross-modal economic reasoning, while promising, has a number of limitations. But data quality, temporal misalignment, modality imbalance [214], market regime shifts and changes in regimes over time (the need for continual

learning), computational complexity affecting performance speed as well as privacy constraints of shared pretrained models and cross-market generalization will remain challenges. Financial market is also subject to shock events or behaviours that might no longer be well-represented in historical datasets. As a result, the absence of any multimodal architecture ought to be understood as an infallible prediction tool.

Future research includes a range of areas including advanced causal reasoning, real-time multimodal learning, multilingual financial intelligence, uncertainty-aware fusion and privacy-preserving financial learning as well as scalable large-scale architecture. But these analytical capabilities could be extended by incorporating other alternative data sources (e.g. supply-chain information, ESG disclosures and macroeconomic events), or even further integration with existing market indicators as inputs in the same framework. Causal cross-modal reasoning holds great potential, in that it could be used to differentiate statistical associations illusory of economic significance from ones which are substantive.

XI. REFERENCES

- [1] Abaci, D. (2019) 'Fin BERT: Financial sentiment analysis with pre-trained language models', *arrive preprint*
- [2] Oval, R., Andrews, N. and Yan, X. (2024) 'Financial forecasting from textual and tabular time series', *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8289–8300.
- [3] Oval, R., Andrews, N. and Yan, X. (2025) 'Multimodal language models for financial forecasting from interleaved sequences of text and time series', *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 987–1001.
- [4] D'Amico, S., Mercuric, F., Nubian, N., Peril, G. and Venter, C. (2026) 'Learning across modalities: A systematic survey of multimodal models for financial analysis', *Information Fusion*, 133, 104249.
- [5] Yang, H., Liu, X.-Y. and Wang, C.D. (2023) 'Fight: Open-source financial large language models', *arrive preprint arXiv:2306.06031*. (arrive)
- [6] Zhang, B., Yang, H. and Liu, X.-Y. (2023) 'Instruct-Fight: Financial sentiment analysis by instruction tuning of general-purpose large language models', *Film Symposium at IJCAI 2023*.
- [7] Zhang, B., Yang, H., Zhou, T., Babar, M.A. and Liu, X.-Y. (2023) 'Enhancing financial sentiment analysis via retrieval augmented large language models', *Proceedings of the ACM International Conference on AI in Finance*.
- [8] Zhang, Y., Dong, Z. and Cu, W. (2026) 'Integrative stock price trend prediction via hierarchical LLM text processing and patch-based transformer with co-attention', *Expert Systems with Applications*, 302, 130441.
- [9] [Authors] (2025) 'MM-transformer: A multimodal approach to economic time series forecasting with textual data', *Applied Sciences*, 15(3), 1241.
- [10] [Authors] (2026) 'Multimodal information fusion for financial forecasting via cross-attention and calibrated uncertainty', *Machine Learning with Applications*, 23, 100840. (Science Direct)
- [11] [Authors] (2026) 'Decoding market dynamics with financial knowledge graphs and cross-modal transformers', *Information Sciences*, article 124031. (Science Direct)
- [12] Name, R., Badness, A. and Talia, M. (2026) 'A hybrid Prophet-based framework for multimodal forecasting with market sentiment signals', *Discover Artificial Intelligence*, 6, 383.
- [13] Aswan, A., Shaker, N., Parma, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 30.
- [14] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT*, pp. 4171–4186.
- [15] Brown, T.B., Mann, B., Ryder, N., Cubbish, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A. et al. (2020) 'Language models are few-shot learners', *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901.
- [16] Dosovitskiy, A., Beyer, L., Kalashnikov, A., Weissenborn, D., Zhao, X., Dehghani, M. et al. (2021) 'An image is worth 16×16 words: Transformers for image recognition at scale', *International Conference on Learning Representations*.
- [17] Tsai, Y.-H.H., Bay, S., Yamada, M., Morenci, L.-P. and Salakhutdinov, R. (2019) 'Multimodal transformer for unaligned multimodal language sequences', *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569.
- [18] Baltrušaitis, T., Abuja, C. and Morenci, L.-P. (2019) 'Multimodal machine learning: A survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423–443.
- [19] Goo, M., Cu, H., Cu, Y., Wang, Y., Wei, R., Ma, J., Zhang, Y. et al. (2023) 'Large language models are zero-shot reasons', *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- [20] Lim, B., Ark, S.Ö., Loaf, N. and Pfizer, T. (2021) 'Temporal fusion transformers for interpretable multi-horizon time series forecasting', *International Journal of Forecasting*, 37(4), pp. 1748–1764.