Transformer-based CoVaR: Systemic Risk in Textual Information

Junyu Chen, Tom Boot, Lingwei Kong
and Weining Wang

Discussion Paper 26/840
16 February 2026

School of Economics
University of Bristol
Priory Road Complex
Bristol
BS8 1TU
United Kingdom

University of BRISTOL
School of Economics

Transformer-based CoVaR:

Systemic Risk in Textual Information

Junyu Chen * Tom Boot * Lingwei Kong * Weining Wang †

February 16, 2026

Abstract

Conditional Value-at-Risk (CoVaR) quantifies systemic financial risk by measuring the loss quantile of one asset, conditional on another asset experiencing distress. We develop a Transformer-based methodology that integrates financial news articles directly with market data to improve CoVaR estimates. Unlike approaches that use predefined sentiment scores, our method incorporates raw text embeddings generated by a large language model (LLM). We prove explicit error bounds for our Transformer CoVaR estimator, showing that accurate CoVaR learning is possible even with small datasets. Using U.S. market returns and Reuters news items from 2006–2013, our out-of-sample results show that textual information impacts the CoVaR forecasts. With better predictive performance, we identify a pronounced negative dip during market stress periods across several equity assets when comparing the Transformer-based CoVaR to both the CoVaR without text and the CoVaR using traditional sentiment measures. Our results show that textual data can be used to effectively model systemic risk without requiring prohibitively large data sets.

Keywords: Systemic Risk, Time-Series Forecasting, Transformer, Large Language Model (LLM), Textual Analysis

JEL Classification Codes: C45, C58, G01

*Department of Economics, Econometrics and Finance, Faculty of Economics and Business, University of Groningen, The Netherlands. Emails: junyu.chen@rug.nl, t.boot@rug.nl, l.kong@rug.nl.

†Corresponding author. School of Economics, University of Bristol, UK. Address: Beacon House, Queens Road, Bristol BS8 1QU, UK. Email: weining.wang@bristol.ac.uk. Phone: +44 7413 466969.

1 Introduction

Between 2007 and 2009, the crisis in the U.S. mortgage market set off a chain reaction that destabilized financial institutions and the global economy. To better anticipate these events, regulators, policymakers, and researchers focused their attention on measuring the systemic risk that arises in a strongly connected economy. While some systemic risk factors can be effectively distilled from numerical market data, additional information can be extracted from textual data, one well-known example being "market sentiment" (Keiber and Samyschew, 2017). However, this text-based risk information is inherently difficult to quantify, and a key open question is how we can effectively extract it from textual sources to obtain better measures and forecasts of systemic risk.

In this paper, we propose a formal econometric framework that uses a Transformer-based architecture (Vaswani et al., 2017) to improve systemic risk analysis by integrating structured financial variables and unstructured textual information such as financial news articles. Our approach proceeds in two steps. First, we employ a large language embedding model that maps variable-length news articles into fixed-size, high-dimensional vectors representing semantic content. Second, we extract the key latent information from these high-dimensional input sequences by leveraging the Transformer's ability to capture nonlinear dependencies and contextual relationships across diverse data sources (i.e., *multimodal data*[1]).

A Transformer is a neural network that relies on the self-attention mechanism to compute latent representations of its input. Traditional neural network architectures (such as convolutional or recurrent networks) rely on fixed parameterized transformations of their inputs, whereas the self-attention mechanism in a Transformer dynamically adapts the transformation based on the input. In other words, self-attention allows the importance of a variable to change based on the surrounding data. For example, a negative news item may be weighted as highly significant for predicting risk when other variables mention "default", whereas the same news item might be down-weighted if the context is a "holiday". By identifying these context-dependent relationships, the Transformer provides the latent representations necessary for the risk prediction task.

A main motivation for using the Transformer is that traditional sentiment scores often miss critical information in textual data or may even misrepresent sentiment indicators relevant for market prediction (Ke et al., 2019; Loughran and McDonald, 2011; Tetlock, 2007). Transformers can effectively learn signals from noisy, high-dimensional textual inputs by down-weighting irrelevant content (Edelman et al., 2022), even when the sample size is limited Mousavi-Hosseini et al. (2025). To show that such results can be also be expected

[1]By multimodal data, we mean inputs of different types, combining structured numerical variables (e.g., returns) with unstructured information such as text.

when estimating the CoVaR on relatively small financial data sets, we establish a new convergence result that complements recent theory on convergence rates for deep neural networks in high dimensions (Schmidt-Hieber, 2020; Suzuki, 2019). Specifically, we show how the quantile regression framework of Padilla et al. (2022) can be extended to incorporate text information through the Transformer architecture. We then show that under the architecture we propose, the convergence rate of the CoVaR estimator only depends logarithmically on the dimension of the input data. As a result, a researcher does not manually need to summarize unstructured text into a low-dimensional object, for example via a sentiment scores, and then specify how these summaries affect risk.

Empirically, we estimate the Conditional Value-at-Risk (CoVaR), a widely used measure of systemic risk, for eight Global Systemically Important Banks (G-SIBs) in the U.S. We consider daily returns between 2006–2013, including both the 2008 financial crisis and the 2011 debt ceiling crisis. We augment the standard numerical data used to estimate CoVaR with a rich set of *Reuters* financial news articles. The Transformer-based method that incorporates the full text data yield more negative CoVaR and ΔCoVaR values during crisis periods than approaches that either exclude news information or reduce it to sentiment scores prior to risk modeling. After addressing potential look-ahead bias, we show that our empirical results remain robust. Furthermore, we demonstrate that incorporating text into the Transformer architecture achieves generally lower prediction loss compared to the model relying only on structured financial return data.

In summary, this paper makes three contributions to advancing the understanding and practical application of multimodal data in systemic risk assessment. First, to our knowledge, we are the first to introduce an econometric framework for analyzing CoVaR using the Transformer-based architecture. Second, we establish upper bounds for its ℓ_2 risk and also generalization error. Finally, we empirically demonstrate that incorporating textual information leads to better predictive performance and alters the CoVaR risk measures. In particular, we identify a pronounced negative dip during the crisis period across several equity assets when comparing the Transformer-based CoVaR to both the CoVaR without text and the CoVaR using the traditional sentiment measure.

Outline. The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the model and estimators. Section 4 presents the main theoretical results. Section 5 reports the findings of the empirical application. Section 6 concludes and discusses directions for future research. Appendix A provides the description of an alternative Transformer architecture. Appendix B discusses the interaction between news content and risk prediction. Appendix C provides additional results from the simulation study. Proofs of all theoretical results are provided in the Supplemental Appendix (SA)

to streamline the main exposition; these results may also be of independent technical interest.

2 Literature Review

Broadly speaking, we can distinguish three approaches to risk modeling in a connected world which we refer to as network-based models, simulation-based approaches and market-based approaches. Early network-based models (Allen and Gale, 2000; Billio et al., 2012; Diebold and Yilmaz, 2014; Acemoglu et al., 2015) highlight the importance of capturing interconnectedness and nonlinear contagion effects. Simulation-based approaches (Gai et al., 2011; Greenwood et al., 2015) underscore amplification mechanisms like fire sales and liquidity spirals, yet they often assume simplified dynamics.

Market-based approaches offer another perspective by deriving systemic risk measures from asset prices and market information: Acharya et al. (2017) proposed Systemic Expected Shortfall (SES) and its variant to estimate capital shortfalls in systemic crises, and Segoviano and Goodhart (2009) developed the Distress Insurance Premium (DIP) using option-implied risk-neutral probabilities. Among market-based approaches, the CoVaR framework proposed by Adrian and Brunnermeier (2016) remains a central tool for quantifying systemic risk. Subsequent contributions, including Hautsch et al. (2014), Härdle et al. (2016), and Keilbar and Wang (2022), extend CoVaR to high-dimensional environments. The high-dimensional setting is essential in applied practice, as systemic risk emerges from the complex interconnectedness of global markets and macroeconomic factors that low-dimensional models fail to address simultaneously. However, these methodologies do not accommodate multimodal data structures, which are increasingly relevant in modern financial systems.

A popular strategy to accommodate information from text is the use of sentiment analysis, where texts are for example classified as positive, negative, or neutral. Early work constructs sentiment indices using dictionary-based word counts: Tetlock (2007) shows that pessimistic media tone predicts short-term market declines, while Loughran and McDonald (2011) refine sentiment lexica to better capture financial context. Later studies, such as Ke et al. (2019), employ topic modeling to assign sentiment weights to terms and aggregate them into document-level scores for return prediction. Even recent neural-network-based studies are sentiment-focused: Schuettler et al. (2024) find that LLMs enhance stock return prediction, Wang and Ma (2024) propose MANA-Net to aggregate sentiment dynamically by market relevance, and Jha et al. (2025) measure popular financial sentiment across languages using LLMs. These findings highlight the value of richer linguistic and contextual signals through market sentiment. However, reducing text to low-

dimensional sentiment categories may overlook other dimensions of information.

Much of the literature has found evidence of other channels affects financial risk and market outcomes beyond sentiment. For example, linguistic characteristics such as grammatical structure, linguistic complexity, etc. affect how information is processed and priced by markets (Loughran and McDonald, 2014; Fedyk, 2024; Kim et al., 2024). In particular, Engelberg and Parsons (2011) and Dougal et al. (2012) establish causal links between differential reporting of events by media or journalists and financial markets. Beyond linguistic characteristics, recent studies demonstrate the informational value of textual data for asset pricing and financial risk (e.g., Adämmer et al. (2025); van Dijk and de Winter (2023); Bybee et al. (2020)), consistent with psychological evidence linking emotions to financial decisions (Lerner and Keltner, 2000, 2001). In addition, topic-based analyses such as Kogan et al. (2009) and Larsen and Thorsrud (2019) show that latent news topics predict volatility and macroeconomic variables. Thus, focusing solely on simplified sentiment analysis may be insufficient to capture the dynamics of risk, particularly during periods of financial stress.

Recent advances in multimodal learning reinforce the view that textual information can provide valuable predictive signals for financial markets beyond traditional sentiment indicators. Transformers and LLMs have been applied to financial prediction using diverse inputs, including earnings-call audio, social media, and news (Li et al., 2020; Zou and Herremans, 2023; Wang et al., 2024; Omranpour et al., 2024; Wang et al., 2025; Yang et al., 2020; Cao et al., 2025). Beyond multimodal prediction, Pele et al. (2026) examine LLMs guided by human prompts for financial forecasting. Our empirical application builds on these approaches by employing a large language embedding model that maps news articles into high-dimensional vectors representing semantic content, rather than collapsing them into a single sentiment score.

3 Methodology

In this section, we first introduce the systemic risk modeling framework and then the Transformer-based architecture used for approximating a quantile function that integrates both financial variables and market-related textual information. In the next section, we formally study its approximation and generalization properties.

3.1 Systemic Risk Modeling

Our CoVaR estimation procedure proceeds in two steps, which builds on the framework of Keilbar and Wang (2022) but modifies the second estimation step. In the first step, we estimate the VaR for each bank using a linear quantile regression model. In the second step, we estimate CoVaR by first fitting a Transformer-based quantile regression conditional on other banks' returns and all financial news, and then plugging in the corresponding VaR estimates from the first step to replace the returns. We do not incorporate text data in the first step to maintain the standard definition of VaR as a quantile of the marginal return distribution. Our focus is therefore on the second step, where we use the Transformer's ability to process textual data.

In the following, we let $R_{j,t} \in \mathbb{R}$ denote the return of bank j at time t, where $j = 1, \ldots, J$ with J fixed and finite, and $t = 1, \ldots, T$. To incorporate textual information, we utilize an embedding mapping, which is referred to in the deep learning literature as an *embedding layer*. This embedding layer maps individual news items into d_e–dimensional numerical vectors (i.e., embedding vectors) that represent their semantic content in a continuous coordinate space. Specifically, let n denote the number of news items observed for bank j at time t. Applying the embedding layer to each news item yields a news-embedding matrix $E_{j,t} \in \mathbb{R}^{d_e \times n}$, where each column represents a single news item and d_e is the embedding dimension. In our empirical application (see Section 5), we employ a *pre-trained*[2] mapping π_e and treat its parameters as fixed. In practice, one may instead *fine-tune*[3] its embedding parameters.

Step 1: Estimation of VaR

A variety of methods has been developed to estimate VaR, including the historical simulation method (Hull and White, 1998), parametric volatility models such as ARCH and GARCH (Engle, 1982; Bollerslev, 1986), approaches based on Extreme Value Theory (McNeil and Frey, 2000), and linear quantile regression (Chao et al., 2015). Any of the approaches described above can be used to construct our VaR estimator. Among these approaches, linear quantile regression provides a transparent, flexible, and easily interpretable alternative (Koenker and Bassett, 1978). It has been widely used in empirical studies linking macro-financial conditions to tail risks (Adrian and Brunnermeier, 2016; Härdle et al., 2016). Following Adrian and Brunnermeier (2016), let $M_{t-1} \in \mathbb{R}^m$ denote a vector of lagged macroeconomic state variables that reflect the general state of the economy. The VaR for bank i at time t is defined as the τ-quantile of its return distribution conditional on M_{t-1}:

$$\mathrm{P}\big(R_{i,t} \leq \mathrm{VaR}^{\tau}_{i,t} \,|M_{t-1}\big) = \tau, \quad \tau \in (0,1). \tag{1}$$

[2] A pre-trained model is trained in advance on large, general data.

[3] Fine-tuning means starting from a pre-trained model and then slightly updating its parameters using task-specific data.

By convention, $\mathrm{VaR}_{i,t}^{\tau}$ is typically negative, reflecting a potential loss. Although alternative sign conventions exist, we follow the one stated above. We employ a linear quantile regression with lagged macroeconomic variables to estimate $\mathrm{VaR}_{i,t}^{\tau}$:

$$R_{i,t} = \alpha_i(\tau) + \gamma_i(\tau)^\top M_{t-1} + \epsilon_{i,t}(\tau) \tag{2}$$

with $\alpha_i(\tau) \in \mathbb{R}$, $\gamma_i(\tau) \in \mathbb{R}^m$. We assume $F^{-1}_{\epsilon_{i,t}(\tau)|M_{t-1}}(\tau) = 0$ with $\epsilon_{i,t}(\tau)$ being the error term for the τ-th quantile. The VaR estimate is then the fitted value of the quantile regression

$$\widehat{\mathrm{VaR}}_{i,t}^{\tau} = \widehat{\alpha}_i(\tau) + \widehat{\gamma}_i(\tau)^\top M_{t-1}. \tag{3}$$

Step 2: Estimation of CoVaR

CoVaR, introduced as a systemic extension of standard VaR by Adrian and Brunnermeier (2016), is defined as the VaR of bank j conditional on a specific state of the financial system. In particular, we consider the specific state as the distress scenario of the financial system, while also taking the market-related textual information (i.e., financial news) into account. For the distress scenario of financial systems, we assume that all other banks are at their VaR level. Formally, we define CoVaR as

$$\mathrm{P}\Big(R_{j,t} \leq \mathrm{CoVaR}_{j,t}^{\tau} \;\Big|\; R_{-j,t} = \mathrm{VaR}_{-j,t}^{\tau},\, E_{j,t}\Big) = \tau, \tag{4}$$

where $R_{-j,t}, \mathrm{VaR}_{-j,t}^{\tau} \in \mathbb{R}^{J-1}$ denote the returns and VaRs of all banks other than j at time period t, respectively, and $E_{j,t}$ relates to the embedded information for j-th bank up to at most time period t (e.g., embedded financial news at time period $t-1$). We estimate $\mathrm{CoVaR}_{j,t}^{\tau}$ using a Transformer-based quantile regression described in Section 3.2. Specifically, the model is given by

$$R_{j,t} = f^*_{j,\tau}\Big(R_{-j,t},\, E_{j,t}\Big) + \nu_{j,t}(\tau), \tag{5}$$

where $f^*_{j,\tau}(\cdot)$ is the true conditional quantile function for bank j, and $\nu_{j,t}$ is an error term satisfying $F^{-1}_{\nu_{j,t}(\tau)|R_{-j,t},E_{j,t}}(\tau) = 0$. The function $f^*_{j,\tau}$ is essential for capturing the complex tail behavior of returns and news. It serves as the channel through which non-sentiment-related risk, as introduced earlier, transmits to the target bank return. Under the distress scenario, we replace $R_{-j,t}$ with the VaR estimates of all other banks $\widehat{\mathrm{VaR}}_{-j,t}^{\tau}$, so that the estimated CoVaR is computed as

$$\widehat{\mathrm{CoVaR}}_{j,t}^{\tau} = \widehat{f}_{j,\tau}\Big(\widehat{\mathrm{VaR}}_{-j,t}^{\tau},\, E_{j,t}\Big), \tag{6}$$

where $\widehat{f}_{j,\tau} \in \mathcal{G}$, with $\mathcal{G}$ denoting a class of Transformer-based neural network functions. The class $\mathcal{G}$, along with the model architecture and estimation procedure, is introduced in the next section.

3.2 Transformer Architecture

We adopt the Transformer architecture for the following reasons. Firstly, a key challenge for our CoVaR framework is integrating numerical and textual inputs without imposing restrictive dependence assumptions. Conventional architectures like CNNs (assuming spatial locality) and recurrent architectures like RNNs (assuming temporal order) (Yin et al., 2017) embed strong functional constraints that may not hold for our data. In contrast, the Transformer architecture enables direct pairwise interactions among all input elements, thereby capturing global dependencies (Vaswani et al., 2017; Cordonnier et al., 2020). This may make the Transformer architecture better suited to settings where neither specific spatial nor temporal structure is clearly known a priori.

Secondly, our application necessitates a model that can learn sparse functions from noisy data. Financial news contains significant redundancy and irrelevance, requiring the model to identify and focus on meaningful information and signals. Additionally, the volume of news varies daily, requiring the use of *padding*[4] to create uniform input sequences. This introduces artificial *tokens*[5] that the model must disregard. The Transformer's self-attention mechanism directly addresses both requirements. It enables the model to dynamically weight input importance, effectively ignoring both irrelevant content and padding tokens. This property aligns with its theoretically established capacity to learn sparse functions ((Edelman et al., 2022)).

Thirdly, another challenge stems from the high-dimensional input, characterized by a large number of news articles n (i.e., the model input length or number of tokens) and a large embedding dimensions d_e. Recent theoretical work suggests this challenge can be mitigated. Trauger and Tewari (2024) demonstrate that, under suitable parameter constraints, the complexity of a Transformer model can even grow independently of the input dimension. In addition, Mousavi-Hosseini et al. (2025) shows that Transformers can have smaller sample complexity that almost independent of the input length, implying superior data efficiency compared to feedforward or recurrent architectures. Inspired by these insights, we adopt a function class composed of Transformer blocks where the parameter norms are constrained. This design ensures that the model's complexity—and consequently, the input dimensions can have little effect on the model complexity,

[4]Padding refers to extending shorter sequences with special meaningless placeholders so that all sequences have the same length. This is necessary because neural networks typically require inputs of uniform size.

[5]In Transformers, a token is the basic unit of text that the model processes. It is analogous to a variable or feature in an econometric dataset. A token is typically represented as a vector. Padding tokens act as placeholders.

as demonstrated in Corollary 1.

The Transformer architecture contains two components that we discuss in depth below: a multi-head self-attention mechanism (MSA) and a position-wise feed-forward network (FFN). We refer to these two components as sublayers. Each sublayer can be wrapped with a residual (skip) connection and layer normalization (He et al., 2016; Ba et al., 2016), which stabilize training and mitigate issues such as vanishing gradients. One MSA sublayer and one FFN sublayer together form a Transformer layer. A Transformer model can contain L such layers stacked sequentially, indexed by $l = 1, \ldots, L$. In the following, we denote the rectified linear unit (ReLU) activation function by $\sigma_R(x) = \max(x,\ 0)$ which operates element-wise, regardless of whether the input is a vector or a matrix. We also define the SoftMax function as

$$\sigma_S : \mathbb{R}^{n\times n} \to [0,1]^{n\times n}, \quad Z \mapsto \left(\frac{e^{Z_{i,j}}}{\sum_k e^{Z_{k,j}}}\right)_{i,j=1}. \tag{7}$$

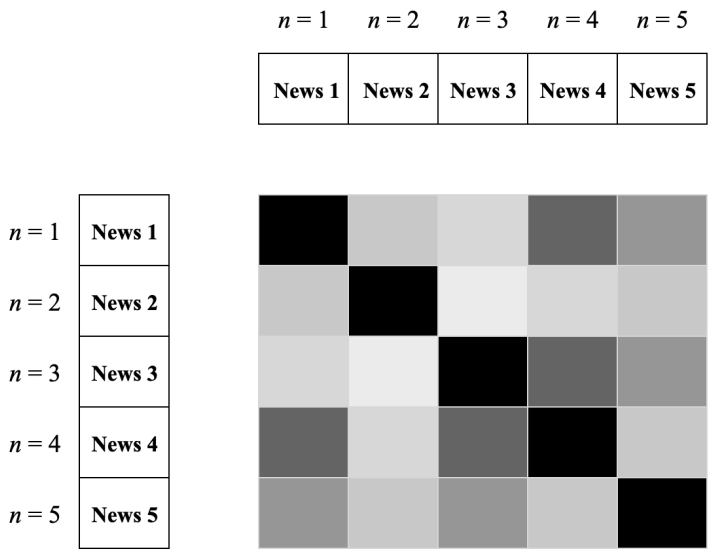
We now zoom in on the Transformer architecture. We consider the Transformer architecture without residual connections and layer normalization in this paper. For a discussion on architectures including these components, please refer to Jiao et al. (2025). While these operations provide additional transformations within each sublayer, they do not alter the fundamental input–output mapping structure of the Transformer block (Edelman et al., 2022). For completeness, we present in Appendix A the standard Transformer architecture, including residual connections and layer normalization.

We take the initial input to our Transformer-based neural network to be a matrix $Z_{j,t} \in \mathbb{R}^{d\times n}$ formed by concatenating the returns $R_{-j,t}$ and the text embeddings $E_{j,t}$. In practice, $R_{-j,t}$ and $E_{j,t}$ may not have compatible dimensions, so we apply a mapping Π to obtain a representation of the desired size. In the empirical application in Section 5, we consider a function Π of the following form:

$$Z_{j,t} := \Pi(R_{-j,t}, E_{j,t}) = \begin{bmatrix} R_{-j,t} \cdot \mathbf{1}_n^\top \\ E_{j,t} \end{bmatrix} \quad \in \mathbb{R}^{d\times n}, \tag{8}$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector of ones and $d = (J-1) + d_e$. This construction replicates the same return vector $R_{-j,t}$ across the n columns and stacks it above each column of $E_{j,t}$. We refer to the resulting columns as *return-augmented news embeddings*.

Figure 1: Example illustration of self-attention scores



Notes: The self-attention mechanism computes pairwise *attention weights* among all elements in the input sequence. In this figure, the attention mechanism learns similarity scores between (return-augmented) news items. Different cell colors correspond to different score values. Darker values indicate stronger interactions.

3.2.1 Self-Attention Sublayer

The self-attention sublayer allows each position (or token) in the input sequence to attend to, and aggregate information from, all other positions in the same sequence (Vaswani et al., 2017). In our application, the input sequence is the concatenated vector of embedded news items and contemporaneous returns from other banks (see Equation (8)). The self-attention mechanism computes pairwise *attention weights* between all elements in this sequence. Intuitively, these weights act as similarity or correlation scores, enabling the model to determine how much focus to place on a given news article when processing another (see Figure 1). A weighted sum of the input values, using these attention weights, produces the sublayer's output, blending relevant information from across the entire input context.

Multi-head self-attention extends the above mentioned idea by splitting the model dimensions d into multiple heads. Let the number of heads be H. We define the l-th multi-head self-attention class $\mathcal{G}_l^{(MSA)}(n, d, H)$ as the set of map $g_l^{(msa)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of the form

$$g_l^{(msa)}(Z) = \sum_{h=1}^{H} W_{l,h}^{(O)} W_{l,h}^{(V)} Z \cdot \sigma_S \left(\frac{Z^\top {W_{l,h}^{(K)}}^\top W_{l,h}^{(Q)} Z}{\sqrt{d}} \right). \tag{9}$$

For each head $h \in \{1, \ldots, H\}$, we denote by $W_{l,h}^{(V)}, W_{l,h}^{(K)}, W_{l,h}^{(Q)} \in \mathbb{R}^{\frac{d}{H}\times d}$, and $W_{l,h}^{(O)} \in \mathbb{R}^{d\times \frac{d}{H}}$ the value, key,

query, and projection matrices, respectively. Both the query and key matrices are learned from data during training as weighted transformations of the input matrix Z. The term $Z^\top W_{l,h}^{(K)\top} W_{l,h}^{(Q)} Z$ captures pairwise interactions between all columns of Z, measuring how similar or relevant one observation (e.g., news item) is to another. Estimating $W_{l,h}^{(K)\top}$ and $W_{l,h}^{(Q)}$ helps us understand which inputs contribute to the risk of bank j. For example, if there is a strong correlation between News 3 and News 4, the corresponding weights will reflect the importance of both components in determining the bank's risk (Figure 1). As another illustration, News 2 and News 3 exhibit weaker interactions relative to other pairs in Figure 1. After dividing by $\sqrt{d}$ and applying the softmax function to normalize the weights, the self-attention mechanism amplifies strong dependencies and filters out weak or noisy signals.

In our risk analysis framework, the input matrix $Z_{j,t,s}$ jointly incorporates numerical return information and textual news content. By concatenating returns with news embeddings, we construct return-augmented news representations. The self-attention mechanism then learns the interactions among these return-augmented news inputs, allowing cross-bank spillovers and textual information to be modeled in a unified manner. In particular, the model identifies which latent information in bank returns and news articles the strongest influence on the risk exposure of firm j. Through this mechanism, the model uncovers dynamic transmission channels through which financial and informational shocks propagate across the system.

3.2.2 Feed-Forward Network Sublayer

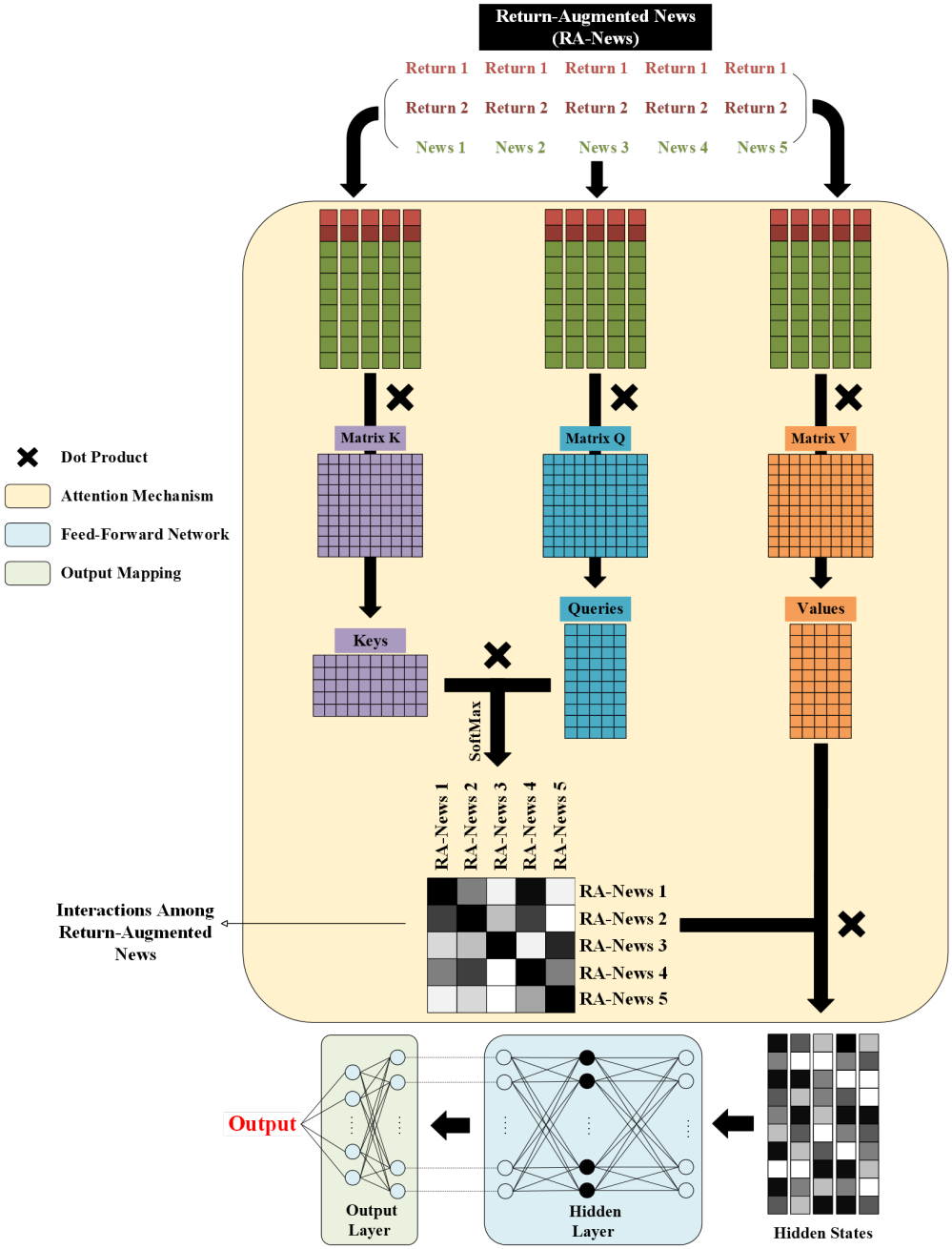
Following the attention sublayer, a feed-forward network (FFN) sublayer applies the same operation to each column (i.e., token) of the input matrix. Its primary role is to provide additional non-linearity and transformation capacity. Recent work by Sonkar and Baraniuk (2023) shows that the FFN prevents the output from collapsing into a narrow subspace of the total embedding space, meaning they become highly correlated and lose their distinctiveness.

The l-th feed-forward network class $\mathcal{G}_l^{(FF)}(d, d_h)$ is defined as the set of maps $g_l^{(ff)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of the form

$$g_l^{(ff)}(Z) = W_l^{(2)}\, \sigma_R\left(W_l^{(1)} \cdot Z + b_l^{(1)}\right) + b_l^{(2)}. \tag{10}$$

where d_h denotes the hidden dimension of the feed-forward network. The parameters are $W_l^{(1)} \in \mathbb{R}^{d_h\times d}$, $b_l^{(1)} \in \mathbb{R}^{d_h}$ and $W_l^{(2)} \in \mathbb{R}^{d\times d_h}$, $b_l^{(2)} \in \mathbb{R}^d$.

Figure 2: Illustration of the Transformer-based CoVaR prediction model



Notes: The model takes numerical inputs (stock returns) and textual inputs (news articles). As an example, consider two bank returns and five news articles as model inputs. News are augmented with return information via concatenation, forming return-augmented news embeddings. The Transformer attention mechanism captures dependencies within the return-augmented news and learns latent representations (hidden states). These hidden states are subsequently passed through a feed-forward network (FFN) and a multilayer perceptron (MLP) to produce a scalar prediction.

3.3 Transformer-based Estimator

We employ a Transformer-based estimator consisting of the previously introduced Transformer architecture followed by a multilayer perception (MLP). To bridge the Transformer output matrix $Z \in \mathbb{R}^{d\times n}$ and the MLP, we apply a linear projection. The linear projection maps the Transformer output onto a vector, which is then passed through the MLP to produce the scalar result. As shown in Figure 2, our estimator first lets the MSA learn the interactions among the return-augmented embeddings, which yields hidden representations that are then passed through the FFN and the MLP to produce the risk prediction.

Formally, we define the Transformer-block function class $\mathcal{G}^{(TF)}(n, d, H, d_h, L)$ as the set of maps $g^{(tf)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ formed by the composition of L Transformer layers. Each layer consists of a multi-head self-attention sublayer and a feed-forward network sublayer:

$$g^{(tf)} = g_L^{(ff)} \circ g_L^{(msa)} \circ \cdots \circ g_1^{(ff)} \circ g_1^{(msa)}. \tag{11}$$

And we define the MLP function class $\mathcal{G}^{(MLP)}(D, \widetilde{d}_m)$ as the set of maps $g^{(mlp)} : \mathbb{R}^d \to \mathbb{R}$ of the form

$$g^{(mlp)}(x) = W_D\sigma_R(\dots \sigma_R(W_1x + b_1)\dots) + b_D, \tag{12}$$

where the dots denote repeated compositions across layers, D is the MLP network depth and $\widetilde{d}_m := \left(d, d_m^{(1)}, \dots, d_m^{(D-1)}, 1\right)$ represents the MLP network width[6]. If all hidden layers have the same width, we denote it by d_m. For $i = 1, \dots, D$, we have W_i as a $d_m^{(i+1)} \times d_m^{(i)}$ weight matrix and b_i as a vector of dimension $d_m^{(i)}$. Thus, the full class of our Transformer-based neural network functions of a given architecture $(n, d, H, d_h, L, D, \widetilde{d}_m)$ takes the following form:

$$\mathcal{G} := \mathcal{G}(n, d, H, d_h, L, D, \widetilde{d}_m) := \left\{g : \mathbb{R}^{d\times n} \to \mathbb{R} \mid g := g^{(mlp)} \circ (g^{(tf)} \cdot w)\right\}. \tag{13}$$

For convenience, we additionally define for a given concatenation function Π (see, e.g., equation (8)),

$$\mathcal{F} := \mathcal{F}(n, d, H, d_h, L, D, \widetilde{d}_m) := \left\{g \circ \Pi : g \in \mathcal{G}\right\}, \tag{14}$$

so that for any $f \in \mathcal{F}(n, d, H, d_h, L, D, \widetilde{d}_m)$, there exists $g \in \mathcal{G}(n, d, H, d_h, L, D, \widetilde{d}_m)$ such that

$$f(R_{-j,t}, E_{j,t}) = g \circ \Pi(R_{-j,t}, E_{j,t}) = g(Z_{j,t}). \tag{15}$$

[6]Depth refers to the number of layers in the MLP, and width indicates the number of neurons within each of those layers.

The Transformer-based quantile estimator is the solution to the following minimization problem for a given class $\mathcal{F}$:

$$\widehat{f}_{j,\tau} = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \rho_\tau \Big(R_{j,t} - f\left(R_{-j,t},\, E_{j,t}\right) \Big), \tag{16}$$

We apply the stochastic gradient descent (SGD) algorithm, see, e.g., (Garg et al., 2022; Bai et al., 2023; Li et al., 2024), for solving the optimization problem in the later applications.

In summary, we construct a Transformer-based quantile regression model for CoVaR that combines peer returns and financial news within a unified architecture (see Figure 2). By explicitly defining the model architecture and its associated function class, we establish a foundation for analyzing both the theoretical properties and empirical performance of the proposed estimator. In the following section, we introduce the consistency property of our CoVaR estimator.

4 Theoretical Results

Notation. We consider the following notation. Let $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$ denote the set of positive real numbers. For vectors, $\|\cdot\|_p$ denotes the ℓ_p norm, and $\|\cdot\|_\infty := \max_{1 \le i \le d} |\cdot|$ is the sup-norm. For a matrix $A \in \mathbb{R}^{a \times b}$, $\|\cdot\|_2$ denotes the spectral norm. Let $\|\cdot\|_{p,q}$ denote the (p,q) matrix norm defined by first taking the p-norm of each row of the matrix, and then taking the q-norm of the resulting vector. We also define $\|A\|_0 = |((i,j) : A_{i,j} \neq 0)|$ and $\|A\|_\infty := \max_{i=1,\ldots,p, j=1,\ldots,q} |A_{i,j}|$. For two sequences, $\{a_n\}_n$ and $\{b_n\}_n$, we write $a_n \lesssim b_n$ if there exists a constant C such that $a_n \le C b_n$ for all n. Moreover, for deterministic sequences $\{a_n\}_n$ and $\{b_n\}_n$ with $b_n > 0$, we write $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $|a_n| \le C b_n$ for all sufficiently large n. For random variables $\{X_n\}_n$ and deterministic $\{b_n\}_n$, we write $X_n = O_p(b_n)$ if X_n / b_n is bounded in probability, and write $X_n \lesssim_P b_n$ if $X_n = O_p(b_n)$.

Define the functional class $V_{(\mathcal{F}, \|\cdot\|_\infty)}(\delta)$ as the logarithm of the δ-covering number of a function class $\mathcal{F}$ under the sup-norm, which measures the complexity of $\mathcal{F}$ by counting the minimum number of balls of radius δ needed to cover it. Specifically, it is defined as $V_{(\mathcal{F}, \|\cdot\|_\infty)}(\delta) := \log N_\infty(\delta, \mathcal{F})$, where N is the smallest number of functions $f_1, \ldots, f_N$ such that every $f \in \mathcal{F}$ is within δ (in sup-norm) of some f_i.

Our main theorem is in the fashion of Suzuki (2019), Schmidt-Hieber (2020) and Padilla et al. (2022) but we adapt these results to the CoVaR framework for Transformers. We first outline the required assumptions.

Assumption 1 (Cumulative Distribution Function Boundedness)**.** *There exists a constant $\kappa > 0$ such that*

for any $\delta_t > 0$ *satisfying* $\max_t |\delta_t|_\infty \leq \kappa$ *we have that, a.s., for all* $j = 1, \ldots, J$,

$$\left|F_{R_{j,t}|x_{j,t}}(f^*_\tau(x_{j,t}) + \delta_t) - F_{R_{j,t}|x_{j,t}}(f^*_\tau(x_{j,t}))\right| \geq \underline{p} \cdot |\delta_t| \tag{17}$$

for $t = 1, \ldots, T$ *and for some constant* $\underline{p} > 0$ *with* $F_{R_{j,t}|x_{j,t}}$ *being cumulative distribution function (CDF) of* $R_{j,t}$ *conditioning on* $x_{j,t}$. *We also require that for all* $j = 1, \ldots, J$,

$$\sup_{z \in \mathbb{R}} p_{R_{j,t}|x_{j,t}}(z) \leq \bar{p}, \qquad a.s. \tag{18}$$

for some constant $\bar{p} > 0$, *where* $p_{R_{j,t}|x_{j,t}}$ *is the probability density function of* $R_{j,t}$ *conditioning on* $x_{j,t}$.

The above CDF Boundedness assumption ensures that the conditional distribution of returns is sufficiently smooth and strictly increasing near the quantile of interest (Padilla et al., 2022). This allows well-defined CoVaR values and ensures that risk estimates respond meaningfully to fluctuations in the underlying data.

Assumption 2 (Lipschitz)**.** *Denote* $\mathcal{E}_j$ *the support of* $E_{j,t}$, *and for all* $j = 1, \ldots, J$, *the target function in (5) satisfies that for all* r_1, r_2 *within the support of* $R_{-j,t}$,

$$\sup_{e \in \mathcal{E}_j} \left|f^*_{j,\tau}(r_1, e) - f^*_{j,\tau}(r_2, e)\right| \lesssim \left\|r_1 - r_2\right\|_1, \tag{19}$$

and for any $f \in \mathcal{F}$ *for the given class* $\mathcal{F}$ *used for fitting, we have that*

$$\sup_{e \in \mathcal{E}_j} \left|f(r_1, e) - f(r_2, e)\right| \leq L_{\mathcal{F}} \left\|r_1 - r_2\right\|_1, \tag{20}$$

with $L_{\mathcal{F}}$ *depending upon the class* $\mathcal{F}$.

The second part in Assumption 2 is trivial satisfied if we impose norm bounds over parameters used in $\mathcal{F}$, as the class $\mathcal{F}$ consists of functions f formed by composing layers that are constructed via linear maps and Lipschitz nonlinear activations/functions. By imposing norm bounds on the parameters of each layer, we ensure each layer is Lipschitz-continuous. Consequently, the overall Lipschitz constant $L_{\mathcal{F}}$ for any $f \in \mathcal{F}$ is bounded by the product of the layer-wise constants. See Appendix E for an explicit example where we consider a one-layer Transformer followed by an MLP with spectral norm bound B.

Assumption 3 (Consistency of VaR)**.** *The estimated VaR is consistent, i.e., for all banks* j,

$$\left\|\widehat{\mathrm{VaR}}^{\tau}_{-j,t} - \mathrm{VaR}^{\tau}_{-j,t}\right\|_1 = O_p(b_T), \tag{21}$$

where b_T is a deterministic sequence such that $b_T \to 0$ as $T \to \infty$.

Under the Lipschitz Continuity and VaR Consistency assumptions, the sampling errors from VaR estimation vanish as the sample size increases, provided the overall Lipschitz constant $L_{\mathcal{F}}$ is of sufficiently small magnitude. This guarantees that the "plug-in" step does not introduce persistent bias, allowing the CoVaR estimator to remain stable and accurate (Patton et al., 2019; Dimitriadis and Hoga, 2025).

Theorem 1 (Out-of-Sample CoVaR Estimator Consistency)**.** *Suppose that Assumptions 1–3 and conditions in Lemma D.1 hold, then the CoVaR prediction satisfies:*

$$\max_{1\le j\le J}\left[\left|\mathrm{CoVaR}^{\tau}_{j,t'} - \widehat{\mathrm{CoVaR}}^{\tau}_{j,t'}\right|\right] \lesssim_P \max_{1\le j\le J} \widetilde{R}_{j,\delta_T,T} + L_{\mathcal{F}}\, b_T, \tag{22}$$

where $\widetilde{R}_{j,\delta_T,T} = \inf_{f\in\mathcal{F}}\left\|f - f^*_{j,\tau}\right\|^2_{\infty} + \left(\delta_T + \sqrt{\frac{V_{(\mathcal{F},\|\cdot\|_\infty)}(\delta_T)}{T}}\right)$ *with* $\delta_T > 0, \delta_T \to 0, \delta_T^2 T \to \infty$ *as* $T \to \infty$, *and* $\widehat{\mathrm{CoVaR}}^{\tau}_{j,t'} = \widehat{f}_{j,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t'}, E_{j,t'})$.

Proof. See Appendix Section D. □

Theorem 1 connects the complexity of the estimator to an upper bound on the generalization error of our estimator. In this sense, it is comparable to Theorem 2.6 in Hayakawa and Suzuki (2020), but is established here under a quantile-loss optimization framework. The result is useful for deriving convergence rates once we impose structure on both the target function and the function class $\mathcal{F}$ used for fitting. Corollary 1 further shows that, under explicit sparsity and smoothness restrictions on the target function, one can obtain a rate whose dependence on the input dimension d and input length n is at most logarithmic.

Although we acknowledge that this rate stated in Corollary 1 can be further sharpened, it already indicates that a plausible approximation remains feasible even with relatively high-dimensional input. We defer the full proof to Appendices G and H, but first briefly summarize the main mechanisms that make such a rate attainable.

In Corollary 1, we follow two core setups similar to the ones employed in Edelman et al. (2022). First, the complexity of the function class $\mathcal{F}$ is governed by the combined capacity of the Transformer blocks and the MLP output layer, and we use norm-based Transformer blocks with sparse MLP, where the $\ell_{2,1}$ norms of the Transformer weights are explicitly bounded. This enables the model complexity to scale only logarithmically in d and n, both of which may be large in practice (e.g., $d = 71$ and $n = 97$ in our application of Section 5),

as captured by the bound

$$V_{(\mathcal{F},\|\cdot\|_\infty)}(\delta) = O\left(d_m \log(d_m) \cdot \log\left(\delta^{-1} D\, d\, d_m^D\right) + \frac{\log(dn)}{\delta^2}\right), \tag{23}$$

where d_m, D are the network width and depth of MLP that possibly grow with T, e.g., in Corollary 1 we choose $D \lesssim \log T$. Second, we assume that the target function has a low-dimensional sparse structure: it depends only on an unknown subset of s relevant input coordinates with s being a finite number, and is β-Hölder smooth. Together, these assumptions imply that self-attention can identify and aggregate the relevant information from the high-dimensional input, while the MLP only needs to learn a smooth mapping on an s-dimensional signal. Hence, the following bias rate is attainable under Assumptions H.1 and H.2 for a given class $\mathcal{F}$ satisfying the aforementioned complexity bound:

$$\inf_{f\in\mathcal{F}} \left\| f - f^*_{j,\tau} \right\|_\infty^2 = O\left(d_m^{-\beta/s}\right). \tag{24}$$

Based on the bias-variance decomposition provided in Theorem 1 and results outlined above, we can achieve the following convergence rate.

Corollary 1 (Convergence Rate)**.** *Suppose that the assumptions of Theorem 1 and Lemma H.1 hold with* $b_T = O(T^{-\alpha})$ *and* $\alpha \geq \frac{2\beta}{s+4\beta}$*, then out-of-sample CoVaR predictor achieves the convergence rate:*

$$\max_{1\leq j\leq J} \left| \mathrm{CoVaR}^\tau_{j,t'} - \widehat{\mathrm{CoVaR}}^\tau_{j,t'} \right| = O_p\left(C_{n,d,T}(T^{-\frac{2\beta}{s+4\beta}} + T^{-1/4})\right), \tag{25}$$

where $C_{n,d,T}$ *is a rate that grows logarithmically with* n*,* d*, and* T*.*

Proof. See Appendix E. □

5 Empirical Study

5.1 Data

We study eight (see Table 1) U.S. global systemically important banks (G-SIBs), as identified by the Financial Stability Board (FSB) with the Basel Committee on Banking Supervision (BCBS) and national regulators. These banks are considered as "too big to fail" because of their role in the global financial system. Our

Table 1: List of U.S. Global Systemically Important Banks

Ticker	Bank Name
GS	Goldman Sachs
WFC	Wells Fargo
JPM	JPMorgan Chase
BAC	Bank of America
C	Citigroup
BK	Bank of New York Mellon
MS	Morgan Stanley
STT	State Street

sample covers daily log returns from October 2006 to November 2013, giving $T = 1776$ observations. The period includes major stress events or market drawdowns such as the 2008 financial crisis, the Goldman Sachs SEC lawsuit in 2010, the Greek debt crisis, and the U.S. debt ceiling disputes of 2011 and 2012. The sample period ensures us the diversity in news topics to study a set of systemic risks.

We use a set of common macro-state variables as risk factors for VaR estimation in (2) (Adrian and Brunnermeier, 2016; Härdle et al., 2016): the Implied Volatility Index (VIX), daily returns on the S&P 500, the spread between Moody's Baa Corporate Bond Yield and the 10-Year Treasury yield, and the yield spread between the 10-Year and 3-Month Treasury yields[7]. For CoVaR estimation, we use both bank stock returns with the financial news articles released from *Reuters*. The news dataset contains 105,359 articles, originally collected by Ding et al. (2014). After removing duplicates, we retain 97,264 articles.

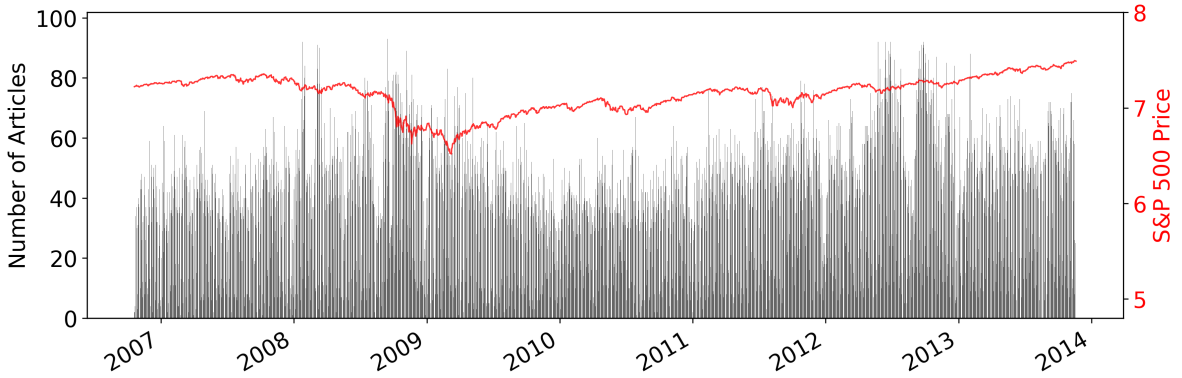
On average, 38 articles are published per day, with daily counts ranging from 1 to 97 and a median of 39. Figure 3 shows that article volume rises sharply during market stress and drawdowns, such as the 2008 crisis and the 2011–2012 U.S. debt ceiling standoff. These spikes coincide with declines in the S&P 500, suggesting that media coverage intensifies in volatile periods. The average article length is 449 words, with a median of 415.

5.2 Data Preprocessing

Preparing the input for our model proceeds in three steps. First, we employ the pre-trained embedding model `gemini-embedding-001` as the embedding mapping π_e to convert n ($n = 97$) textual news used at time t into numerical news embeddings $E_{j,t} \in \mathbb{R}^{d_e \times n}$ (we choose $d_e = 64$). For days on which n is smaller than 97, we apply padding to the news embeddings (see the explanation of padding in the footnote in Section 3.2). We

[7]VIX, S&P 500 returns, and stock prices are from Yahoo Finance. Credit spreads and yield spreads are from the Federal Reserve Bank of St. Louis (FRED).

Figure 3: Daily news article frequency and stock market index



Notes: The left y-axis shows the daily frequency of *Reuters* financial news articles. The right y-axis shows the logarithm of the S&P 500 price index.

consider all news articles observed up to time t within a five-day look-back window. This choice is motivated by Garcia (2013), who documents that market reactions to news typically reverse within four trading days, suggesting that such an effect is short-lived and largely driven by non-fundamental factors. Second, because the embeddings do not inherently have temporal order, we inject positional information to preserve the chronological information of the news sequence. Following Vaswani et al. (2017), we add fixed sinusoidal positional encodings (PE)[8] to the sequence $E_{j,t}$. This preserves the temporal sequence information of news articles. Third, we augment the news embeddings with numerical return information by concatenating the returns of the remaining $J-1$ ($J=8$) banks with $E_{j,t}$, yielding the return-augmented news embeddings $Z_{j,t} = \Pi(R_{-j,t}, E_{j,t})$. We then take $Z_{j,t}$ as our model input, which consists of 97 input tokens, each with 71 (i.e., 64 plus 7) dimensions.

As mentioned above, we use `gemini-embedding-001`, a pre-trained embedding model released in June 2025 by *GoogleAI*. This model generates dense vectors of 3072 dimensions. It takes a maximum of 2048 tokens per input, which corresponds to about 1228–1638 English words[9]. The majority of articles in our dataset are shorter than this limit, with only about 1% exceeding it (see Figure 4). The longest article has 9471 words.

Pre-trained embeddings are known to be effective general-purpose semantic representations (Zhao et al., 2025) and have shown good performance in various natural language processing tasks such as semantic search,

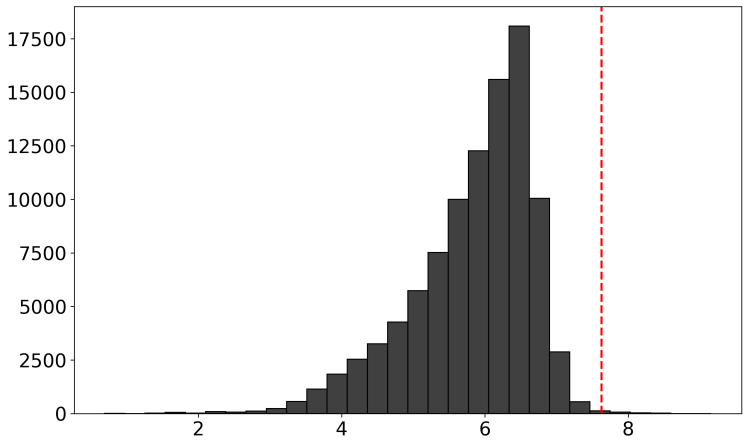
[8]The positional encoding for position k and dimension i is defined as:

$$PE(k,2i) = \sin\left(\frac{k}{10000^{2i/d_e}}\right),$$
$$PE(k,2i+1) = \cos\left(\frac{k}{10000^{2i/d_e}}\right),$$

where k is the temporal position within the 5-day look-back window. These encodings are scaled by the norm of the input embeddings.

[9]According to GoogleAI documentation, 100 tokens are about 60–80 English words.

Figure 4: Distribution of article lengths



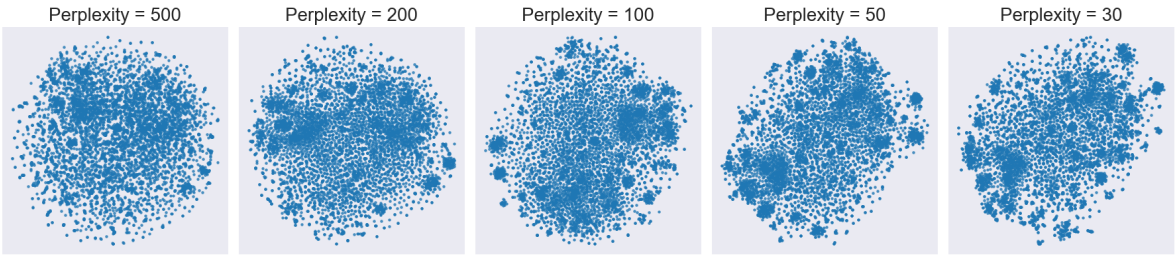
Notes: Article length is measured in number of words. The x-axis shows the logarithm of article length, and the y-axis shows frequency. The dashed red line indicates the 1228 word cutoff, corresponding to 2048 tokens.

text classification, and clustering (Lee et al., 2025). An important property of `gemini-embedding-001` is that the embedding dimensions are ordered by importance, with the most informative content appearing in the first dimensions. This property enables dimensionality reduction by keeping only the top leading dimensions of each vector embedding (Kusupati et al., 2024; Kim et al., 2024). Following this idea, we keep the top 64 dimensions of each embedding, since using all dimensions would make the model too large and require much more training data. We can then represent a collection of n articles by a semantic embedding matrix with size $\mathbb{R}^{d_e \times n}$, where each row is a $d_e = 64$ dimensional embedding. This matrix is then used as the input in the downstream risk prediction task.

To validate that the embeddings capture meaningful semantic patterns in the text data, we study their spatial structure in the embedding space. Since the embedding model is designed to map semantically similar texts to nearby points in the vector space, we expect that texts with similar content will produce embeddings with smaller distances (e.g. Euclidean, cosine distances etc.). For example, the sentences "The market is going to crash" and "The market is expected to plunge" should yield embedding vectors that are close to each other, reflecting their semantic similarity.

First, we apply the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to project the high-dimensional embeddings into two dimensions for visualization purposes. t-SNE is one of the most popular methods for revealing intrinsic structures in high-dimensional datasets, including trends and patterns, through a nonlinear dimensionality reduction technique (Cai and Ma, 2022). This is useful for high-dimensional data that lie on several different but related low-dimensional manifolds (van der Maaten and Hinton, 2008). For example, news items can report various sub-events or highlight different facets within

Figure 5: t-SNE visualization of Goldman Sachs–related news articles



Notes: The figure shows t-SNE embeddings of 6,124 news articles related to Goldman Sachs under different perplexity parameter settings.

the same topic.

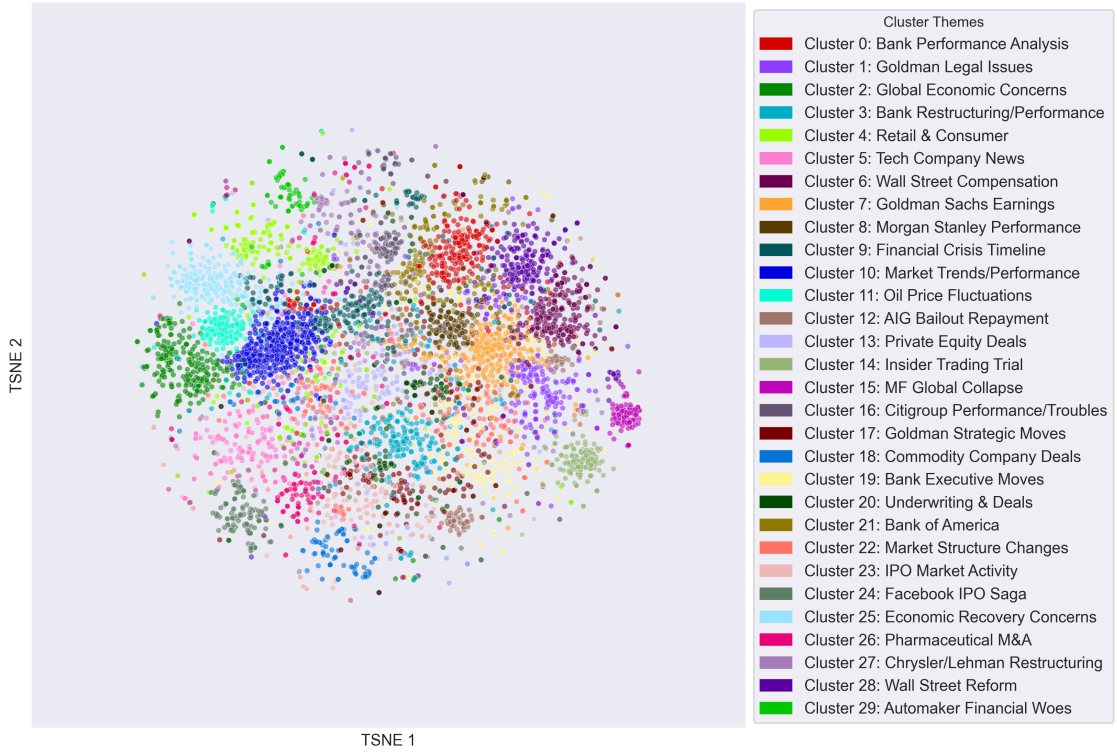
t-SNE includes a parameter called "perplexity", which balances the focus between local and global structures among data points. Local structure refers to the relationships between each point and its nearest neighbors, capturing fine-grained similarities. Global structure refers to the broader arrangement of clusters relative to one another. Lower perplexity values emphasize local relationships, producing smaller, tighter clusters, whereas higher perplexity values emphasize global structure, resulting in larger, more diffuse clusters. As shown in Figure 5, clusters become smaller and more concentrated with lower perplexity values, and larger and more diffuse when the algorithm emphasizes global structure. However, in all cases, the visualizations reveal a pattern of clustering, suggesting that the embeddings form coherent groups in the reduced space.

Next, we apply the KMeans clustering algorithm directly in the original high-dimensional embedding space. Since KMeans is an unsupervised learning algorithm, the number of clusters must be specified in advance. Choosing a smaller number of clusters results in broader, more general themes, while a larger number yields more granular clusters with narrower topics. As an example, we examine all 6124 news articles related to Goldman Sachs[10]. Figure 6 shows the result of applying KMeans with 30 clusters to these articles. For each cluster, we use the LLM `gemini-2.0-flash` to assign a common theme to the 15 most representative articles. These representative articles are defined as those closest to the cluster centroid. The LLM generates a thematic summary based on the topic shared among the 15 selected articles.

Taken together, the t-SNE visualizations and the KMeans clustering results provide qualitative evidence that the embeddings capture meaningful semantic structure in the news corpus. The presence of stable clustering patterns across different perplexity settings, along with the emergence of interpretable themes within clusters in the original embedding space, suggests that semantically related articles are mapped to nearby regions.

[10]These articles are selected based on whether the title, summary, or body text contains the phrase "Goldman Sachs."

Figure 6: KMeans clustering on Goldman Sachs–related news (30 clusters)



Notes: The clustering is visualized using the t-SNE algorithm, and each cluster represents a group of semantically similar news articles.

5.3 CoVaR Prediction

We input the high-dimensional return-augmented news embeddings into our Transformer-based model with one Transformer-block layer and a two-layer MLP. We set the hidden dimensions of both the FFN and the MLP to $d_h = d_m = 64$. We estimate CoVaR in two steps as described in Section 3. First, we estimate VaR for all banks using linear quantile regression on macro-state variables, following Adrian and Brunnermeier (2016) and Härdle et al. (2016). Second, we estimate CoVaR for our target bank using a Transformer-based model that incorporates both numerical (i.e., stock returns) and textual information (i.e., financial news articles).

Our primary objective is to evaluate the added value of incorporating textual information in the second step of CoVaR framework. The dependent variable is the log of the stock return of the target bank. The independent variables include the contemporaneous log stock returns of other banks and news embeddings from the past five days.

To evaluate out-of-sample performance, we split the dataset into three subsets: training ($\mathcal{D}_1$), validation

($\mathcal{D}_2$), and test ($\mathcal{D}_3$). The training set covers 40% of the full dataset and is used to fit the model. The validation set takes 20% and is used for regularization (i.e., early stopping) and hyperparameter tuning. We use the remaining 40% as the test set to evaluate out-of-sample predictive performance of our Transformer. Each subset contains at least one major stock market downturn. We highlight the first and third big market downturns with gray-shaded areas in the following figures and skip the second one, since the second occurs in the validation set and is not of interest to us. The first downturn corresponds to the 2008 financial crisis, which is determined by the NBER (National Bureau of Economic Research) recession period. The third downturn corresponds to the 2011 U.S. debt ceiling crisis, which we determine using quarterly windows. Note that we train the model only once and then use the trained model to perform one-step-ahead predictions on the test data in a rolling-window manner. This means that for each test period, we use all available information up to time t to predict the outcome at time $t+1$.

We train our Transformer-based model using SGD with mini-batches (Marek et al., 2025). In the hyperparameter search, we experiment with three learning rates (0.00015, 0.0015, and 0.015) and three batch sizes (32, 64, 128). We then select the optimal hyperparameters that achieve the lowest loss on the validation data. We choose 200 training epochs with early stopping when the loss does not decrease for 50 consecutive epochs. Because our Transformer incorporates textual information, the optimization becomes much more difficult, and the training objective may include multiple local minima. We acknowledge that some training runs may not converge to the same solution.

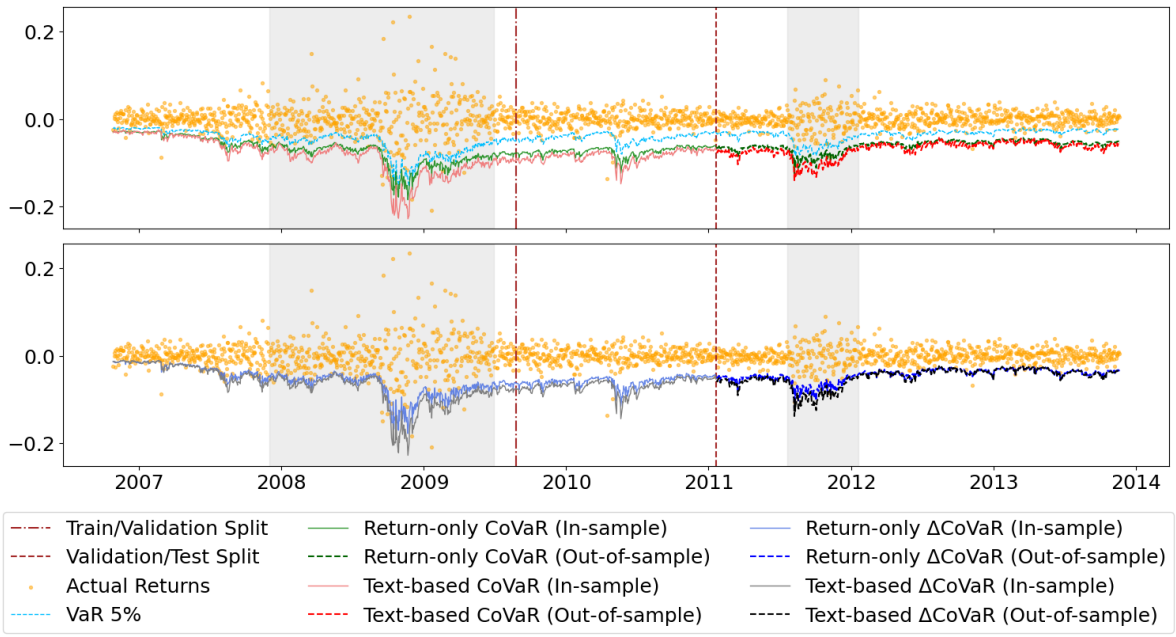
Figure 7 shows the CoVaR and ΔCoVaR estimates for Goldman Sachs, with and without textual information (see Figure 15 in Appendix I for other assets). ΔCoVaR is an alternative risk measure proposed by Adrian and Brunnermeier (2016) that builds upon the CoVaR framework. It quantifies the spillover effect by measuring the change in the system's risk (CoVaR) when a specific bank j shifts from its median state to a state of financial distress. We denote ΔCoVaR as

$$\Delta\text{CoVaR}^{\tau}_{j,t} = \text{CoVaR}^{\tau}_{j,t}\left(\text{VaR}^{\tau}_{-j,t}\right) - \text{CoVaR}^{\tau}_{j,t}\left(\text{VaR}^{0.5}_{-j,t}\right). \tag{26}$$

Importantly in this context, the τ used for the conditional systemic risk (CoVaR) and the τ defining the banks' distress levels (VaR) can be distinct.

The text-based CoVaR is estimated using our Transformer-based model introduced in Section 3.2, and the non-text-based (i.e., return-only) estimates are computed from a two-layer MLP neural network with hidden dimensions 64. During periods of non-crisis and normal market activity, the gap between VaR and CoVaR tends to widen, reflecting the silent accumulation of systemic risk. This supports the perspective

Figure 7: VaR, CoVaR, and ΔCoVaR estimates for Goldman Sachs at $\tau = 5\%$



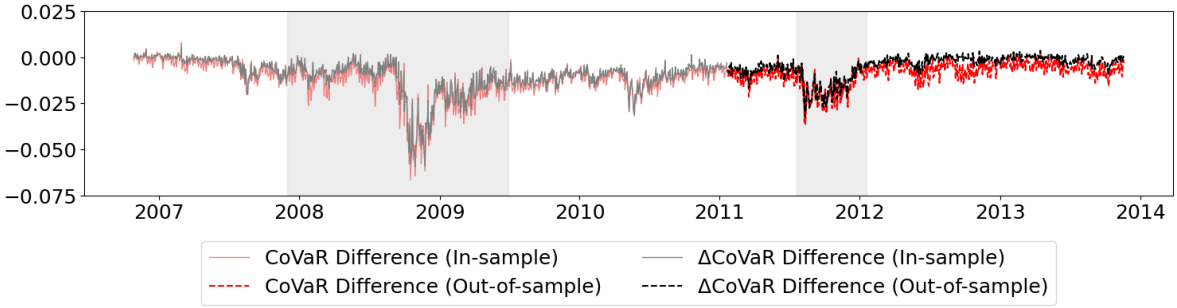
Notes: The estimates are shown with and without textual information. The time series is divided into three parts: the training period (in-sample), the validation period (in-sample), and the test period (out-of-sample), separated by red lines. The **top panel** shows returns, VaR, and CoVaR, while the **bottom panel** shows ΔCoVaR. Grey areas indicate major market downturns.

of Adrian and Brunnermeier (2016), who argue that systemic risk accumulates silently during calm periods and materializes during crises. The latter is evident during market downturns, where the gap narrows as individual institutions approach their VaR thresholds, reflecting the realization of systemic risk.

Comparing the Transformer-based CoVaR model with text against the baseline MLP without text shows that incorporating financial news leads to noticeably different risk estimates during periods of market stress (See Figure 8). The largest absolute differences appear during the 2008 crisis, especially in October, when the differences reach about 0.06–0.065 for both CoVaR and ΔCoVaR. At that time, the global markets were highly unstable and concern about the viability of major financial institutions intensified.

During the 2008 crisis, the media narrative was dominated by panic about the global banking system, fears of a deep recession, and widespread uncertainty about whether government interventions would be sufficient to stabilize financial markets. For example, *Reuters* reported news such as "Germany slashes 2009 growth forecast as crisis bites", "Recession looms despite global interventions", and "Swiss banks raise emergency funds to fight crisis". These news articles carry signals about systemic fragility, which push the model toward more negative CoVaR and ΔCoVaR estimates compared with the model that relies purely on numerical variables. Other late-2008 dates (e.g., November, December) show similar gaps. In relative terms, the two models differ by 40–50% during this crisis.

Figure 8: Differences in CoVaR and ΔCoVaR estimates for Goldman Sachs



Notes: Differences are computed as the text-based Transformer estimates minus the return-only MLP estimates.

Another significant drop shows up in mid-2010. The news then focused on the euro-zone debt crisis and systemic risk in Europe. For example: "EU urges Greece to stick to austerity, pension plan", "Euro-zone troubles may roil stocks", and reports on flash-trading risks such as "Naked swaps, flash trading and other key terms." This information again increase perceived risk, and the text-based model reflects that in more negative CoVaR estimates.

For the out-of-sample period, the largest absolute differences occur during the volatility episode from July 2011 to January 2012. The absolute gaps reach about 0.033–0.037 for CoVaR and ΔCoVaR, with relative deviations of around 33% to 40% for ΔCoVaR. The dominant news themes during this time concern the U.S. debt-ceiling crisis, such as "United States loses prized AAA credit rating from S&P" and "Obama officials attack S&P's credibility after downgrade", as well as the euro-zone sovereign crisis, such as "Euro zone investors gloomiest for nearly 2 years" and "ICE bars MF Global from floor, customers can unwind only"[11].

Across both crisis periods (and also the one in 2010 included in the validation set), the sign of the differences is consistently negative, showing that the model with text produces more negative CoVaR and ΔCoVaR estimates than the model without text. In other words, during crisis episodes, incorporating financial news systematically amplifies the perceived downside tail risk, particularly during crisis episodes. We show an extra news content analysis in Appendix B to better understand why the inclusion of news embeddings leads to more negative CoVaR predictions during periods of financial distress.

[11]Based on *Reuters* news reports, the collapse of MF Global in 2011 was caused by risky 6 billion euro bets on European sovereign bonds, heavy leverage, and the alleged misuse of client funds to cover losses, culminating in bankruptcy.

Table 2: Cumulative three-month average losses for Goldman Sachs

	Text-based Transformer	**Return-only MLP**
3 months	**0.10**	0.13
6 months	**0.11**	0.14
9 months	**0.12**	0.15
12 months	**0.12**	0.15
15 months	**0.12**	0.16
18 months	**0.12**	0.15
21 months	**0.12**	0.15
24 months	**0.12**	0.15
27 months	**0.12**	0.15
30 months	**0.12**	0.15
33 months	**0.11**	0.15
Full Test Period	**0.11**	0.15

Notes: The table reports cumulative three-month average losses of the test data for the text-based Transformer and return-only MLP CoVaR models over twelve quarterly evaluation periods for Goldman Sachs. Entries are reported in units of 10^{-2}, (i.e., values are multiplied by 100).

5.4 Average Quantile Loss

We show that our CoVaR measure can be accurately estimated using a residual-based analysis. In particular, we evaluate the average quantile (AVQ) loss on the test data for CoVaR models with and without textual information. Starting from January 21, 2011 (the beginning of the test period), we compute quarterly average losses and extend the evaluation cumulatively every three months.

Table 2 reports the loss comparison results for Goldman Sachs. Over the full test period, the text-based CoVaR model exhibits slightly lower losses than the CoVaR model without text. However, we observe a noticeable increase in losses between the 9th and 15th months for both models. These periods cover exactly a major market downturn: on August 5, 2011, Standard & Poor's downgraded U.S. sovereign debt from AAA to AA+ for the first time in history; and later, until December 2011, the U.S. debt limit was raised three times after Congress failed to block the increases or pass a Balanced Budget Amendment[12]. *Reuters* news articles during this period documented the intense political conflict between Congress and the Obama administration over raising the debt ceiling, raising fears of a potential default. The textual data from this time contains information relevant to systemic stress: terms such as "default risk," "debt ceiling," and "downgrade" become frequent in financial news. The Transformer-based model takes this textual information as input and achieves a smaller increase in loss during this crisis period.

We report the AVQ losses for all the banks in Table 3. Overall, the two models show broadly similar

[12]See details in Congressional Research Service *The Debt Limit: History and Recent Increases*, and U.S. Government Accountability Office *Debt Limit: Analysis of 2011–2012 Actions*

Table 3: Average quantile losses of CoVaR models for eight major banks

Ticker	Test Loss in First 12 Months		Total Test Loss	
	Transformer	**MLP**	**Transformer**	**MLP**
GS	**0.12**	0.15	**0.11**	0.15
WFC	**0.16**	0.16	0.14	**0.12**
JPM	**0.11**	0.14	**0.12**	0.14
BAC	**0.19**	0.19	**0.16**	0.16
C	**0.19**	0.20	**0.18**	0.18
BK	**0.14**	0.14	0.12	**0.12**
MS	0.21	**0.18**	0.18	**0.14**
STT	**0.15**	0.15	0.15	**0.15**

Notes: The table reports average quantile losses of CoVaR models with and without textual information. Test losses are shown for the first 12 months (crisis period) and for the total period. Bold numbers indicate the lower loss between the Transformer and MLP models for each bank. Entries are reported in units of 10^{-2}, (i.e., values are multiplied by 100).

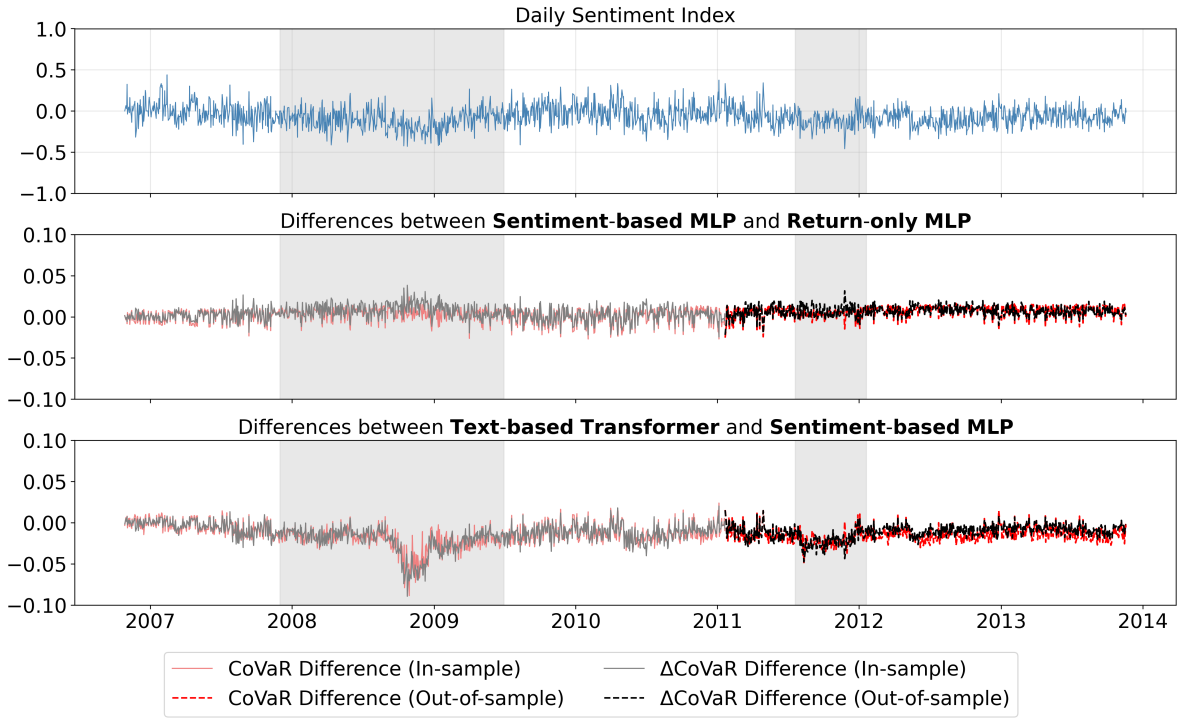
total losses on test data. The inclusion of the financial news does not dramatically change overall predictive accuracy across the full test sample. The performance differences between the two models are generally small across the eight assets. However, we observe two things: 1) the differences between the text-based CoVaR and the CoVaR without text consistently display a negative dip for all banks during the crisis, regardless of whether the total test losses are higher or lower for either model (See Figure 15); and 2) most losses in the first 12 months, which include the major market downturn, tend to be lower or at least comparable for the Transformer-based model with text.

The generally similar total losses suggest that news content may be noisier and provide limited incremental information for risk prediction during stable periods, which may offset the benefits of text integration in evaluations. However, when we focus on the first 12 months, which cover the crisis period, the lower losses observed during this time indicate that textual information becomes more valuable in more uncertain market environments.

5.5 Comparison with Sentiment-Based Measures

We compare our results with sentiment-based CoVaR estimates. Specifically, we use `FinBERT`, a financial-domain pre-trained language model, to classify each news article into one of three sentiment categories: positive, negative, or neutral. These sentiment information are then incorporated into two alternative modeling frameworks: MLP and Transformer. We estimate CoVaR using each sentiment-based model and compare the resulting MLP- and Transformer-based sentiment CoVaR estimates with our text-based CoVaR

Figure 9: Differences in CoVaR and ΔCoVaR for Goldman Sachs across models



Notes: The **top panel** shows the daily sentiment index over time. The **middle panel** shows differences between the sentiment-based MLP estimates and return-only MLP estimates. The **bottom panel** shows differences between the text-based Transformer estimates and the sentiment-based MLP estimates. Red lines indicate differences in CoVaR, and black lines indicate differences in ΔCoVaR.

estimates.

For the first method, to let the MLP take sentiment as an input variable, we construct the daily sentiment index following a commonly used method in the literature (Tetlock, 2007; Loughran and McDonald, 2011; Nassirtoussi et al., 2014):

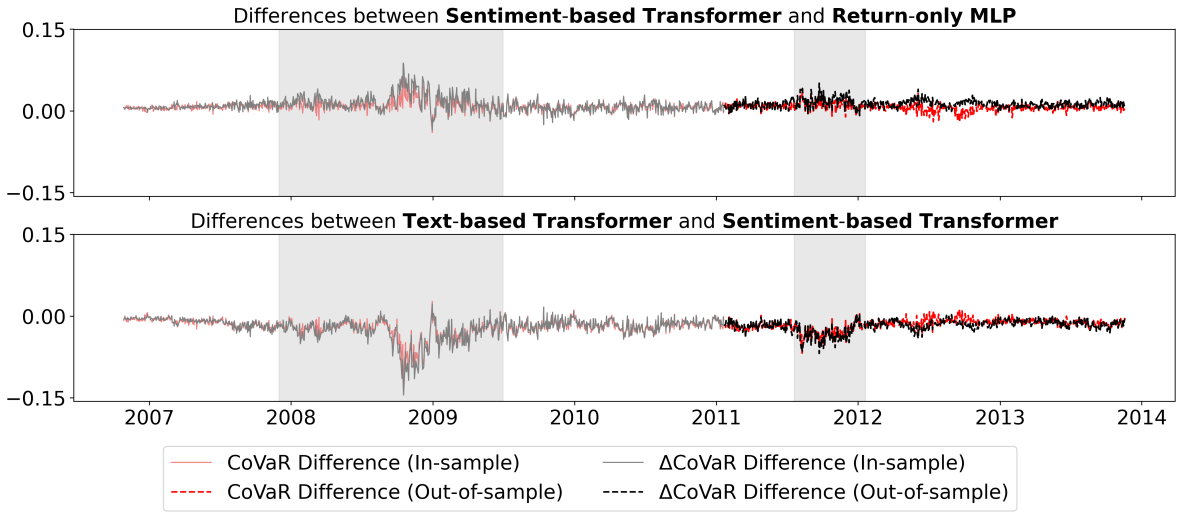
$$\text{Sentiment Index}_t = \frac{\#\text{ Positive}_t - \#\text{Negative}_t}{\#\text{ Positive}_t + \#\text{ Neutral}_t + \#\text{ Negative}_t}, \tag{27}$$

where "#" denotes the number of articles in each sentiment category on day t. The index reflects the net sentiment (positive minus negative) scaled by the total number of articles, representing the overall sentiment of financial news on that day (see the top panel of Figure 9). The observed sentiment index fluctuates around zero and lies within the range $[-0.5, +0.5]$. Negative sentiment occurs during crisis periods, but the magnitude is small. We augment the return-only MLP model by adding the sentiment index as an additional input variable. As shown in the middle panel of Figure 9, the sentiment-based CoVaR adds limited value to systemic risk estimation compared to the return-only MLP CoVaR model. We provide some

potential explanations. First, sentiment is only one aspect of human language and cannot fully capture the complexity of language. Relying only on sentiment leads to the loss of important information relevant to financial markets. Second, even if we assume sentiment is the only key information in textual data, investor sentiment still varies significantly across contexts and is not uniformly informative (Wang and Ma, 2024). Third, common aggregation methods such as averaging or polarity counts are typically static and equal-weighted. They treat all news sentiments as equally important through the time. This overlooks time-varying importance in news information. These aggregation processes lead to the homogenization of sentiment index (Wang and Ma, 2024). This homogenization can be both cross-sectional and temporal: (1) The sentiment scores tend to converge toward the mean and thereby smoothing out the distinctive crucial information in individual news items. The extreme sentiments are diluted, causing predictive information loss; (2) The similarity in average scores across different days homogenizes daily news sentiments, making it difficult for models to capture meaningful time-series variation in investor sentiment. The homogenization mechanism helps explain the small magnitude of day-to-day sentiment fluctuations shown in the top panel of Figure 9. Moreover, news and prices may have interactions in real world, which the common sentiment-based approaches fail to capture.

The second method we use computes a sentiment-based Transformer CoVaR. Instead of feeding high-dimensional text embeddings, we map each document simply as one of three numerical sentiment labels (e.g., 1, 2, 3) to represent negative, neutral, and positive. Then we apply the same procedure as the text-based Transformer to construct the input matrix. We obtain the same conclusion as using the first sentiment-based MLP model: the sentiment-based CoVaR adds limited value to systemic risk estimation compared to the return-only MLP CoVaR model. The largest differences between the two models occur during the 2008 financial crisis. We zoom in and check the sentiment classifications of the news. We zoom in and check the sentiment classifications of the news. Indeed, some news are classified as negative when they are extremely negative, such as "AIG struggles to survive financial tsunami". However, there are also news classified as positive, such as "AIG gets New York's help in accessing $20 billion". But is this really a positive news? One can argue that AIG receives governmental support, but on the other hand, when an institution really needs government intervention, it also means the situation is already extremely bad. In addition, "Chrysler sees U.S. auto market gains toward end 2009" receives a positive label despite the first half of the article describing severe declines in U.S. auto sales, such as *"... auto sales were down more than 11 percent through the first eight months ..."*. Classifying articles into only three categories will ignore the information in the first part. Moreover, more than 50% news during crisis are simply classified as neutral. Thus, collapsing articles into three sentiment categories discards the nuances contained in mixed-tone narratives.

Figure 10: Differences in CoVaR and ΔCoVaR for Goldman Sachs across models



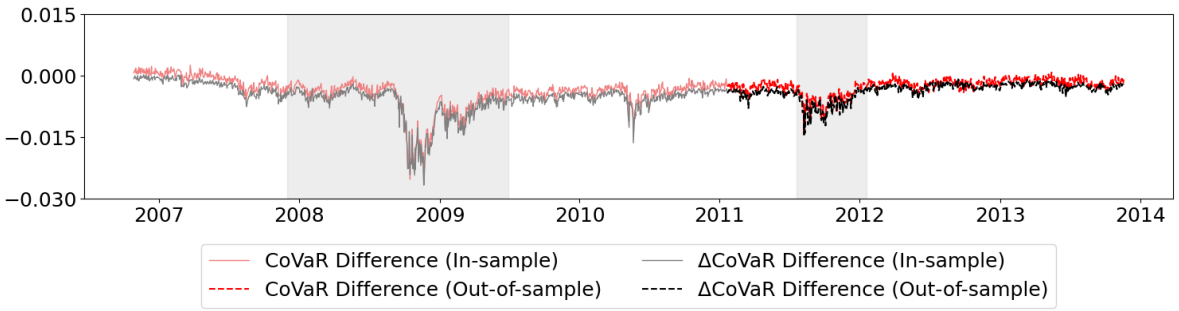
Notes: The **top panel** shows differences between the sentiment-based Transformer estimates and the return-only MLP estimates. The **bottom panel** shows differences between the text-based Transformer estimates and the sentiment-based Transformer estimates.

The limitations of the sentiment-based CoVaRs become obvious, when comparing them with the text-based Transformer CoVaR estimates (i.e. the one that uses text embedding). The text embeddings with richer representations that capture deeper linguistic and contextual information, allowing them to detect more relevant information in news, especially during periods of market stress. As shown in the bottom panels of Figures 9 and 10, the differences between the sentiment-based models and the text-based Transformer are most obvious during crisis periods, where negative dips in CoVaR estimates reappear. This is similar to the differences observed between text-based and return-only CoVaR estimates. This shows that sentiment-based CoVaR modeling overlook important information present in the text, which on the contrary can be captured by using text embeddings.

5.6 Look-Ahead Bias

Applying the *GoogleAI* embedding model to transform news articles into embedding vectors may introduce a look-ahead bias. The `gemini-embedding-001` model may have been trained on data that include future information beyond the analysis period. For instance, if the training data covers the 2008 financial crisis and subsequent bank failures, the model may produce highly similar embeddings for news about different banks. This can be problematic if it occurs even for articles published before the crisis, when news mentioned banks were operating normally. Such bias may skew the embeddings, which overlook subtle but important distinctions in news across different financial institutions.

Figure 11: Differences in CoVaR and ΔCoVaR for Goldman Sachs: with vs. without masking



Notes: Differences are computed by subtracting the CoVaR with masking from the CoVaR without masking. Red lines show differences in CoVaR, and black lines show differences in ΔCoVaR. Negative values indicate that the model without masking produces more negative predictions. Shaded regions correspond to identified crisis periods, and dashed vertical lines mark key financial stress events.

Mitigating the look-ahead bias is challenging. Ideally, we would use a language model trained only on data preceding our evaluation period. However, since the LLM boom began after 2020, it is difficult to find a model that meets the data requirements of our application and contains sufficient coverage of crisis periods for analysis. Developing a new LLM would be computationally and resource intensive. In addition, *GoogleAI* does not disclose the knowledge cutoff date of `gemini-embedding-001`. One dataset used in its training is *FRet*, which includes diverse topics such as blog posts, news articles, Wikipedia-like content, and forum discussions (Lee et al., 2024). Therefore, there is a high potential for look-ahead bias.

We address this bias by masking the names of the eight banks and other identifying information to prevent the model from exploiting bank-specific cues. Similar to Breitung and Müller (2025) and Glasserman and Lin (2023), we apply a Named Entity Recognition (NER) model from the Python package *spaCy* to detect the identifying information of banks. Breitung and Müller (2025) empirically demonstrate that using this advanced model ensures high masking accuracy.

We find that masking or not masking identifying information does not change the overall conclusion: the CoVaR still becomes more negative during crisis periods when financial news is incorporated. However, there are small numerical differences remain between the models with and without masking (see Figure 11). Negative values in the figure indicate that the model without masking produces more negative CoVaR estimates. The largest absolute differences occur between October and November 2008, when the CoVaR and ΔCoVaR gaps reach about 0.022–0.027. In relative terms, the Transformer-based model without using masking deviates from the Transformer-based model using masking by roughly 12–14% for both CoVaR and ΔCoVaR. Over the out-of-sample period, the largest gap appears from August to October 2011, with absolute differences of about 0.015 for CoVaR and ΔCoVaR (roughly 9–12% in relative terms). A likely

explanation is that masking reduces how often multiple banks appear in the same article, which weakens the model's ability to infer relationships between institutions. This may affect the systemic risk estimates.

6 Conclusion

We develop a Transformer-based framework for systemic risk modeling that integrates both structured financial data and unstructured textual data. Our approach extends the CoVaR methodology to incorporate high-dimensional, multimodal inputs. We provide empirical evidence of improved tail-risk estimation during periods of financial stress. Specifically, we show that including textual data from financial news results in more negative CoVaR and ΔCoVaR values during crises under a better predictive performance, highlighting the importance of market-related information contained in textual data. In addition, forward-CoVaR, as another systematic risk measure used for early warning, can be interesting for future empirical work.

We establish an upper bound on the ℓ_2 risk of the Transformer-based quantile regression. In particular, we also provide the convergence rate of our procedure for an explicit case, demonstrating that it can efficiently learn complex nonlinear dependencies in high-dimensional settings. This complements existing literature on deep learning for financial risk modeling and provides a rigorous econometric foundation for integrating Transformers into systemic risk analysis.

References

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2015). Systemic risk and stability in financial networks. *American Economic Review 105*(2), 564–608.

Acharya, V. V., R. F. Engle, and M. Richardson (2017). Measuring systemic risk. *Review of Financial Studies 30*(1), 2–47.

Adrian, T. and M. K. Brunnermeier (2016). CoVaR. *American Economic Review 106*(7), 1705–1741.

Adämmer, P., J. Prüser, and R. A. Schüssler (2025). Forecasting macroeconomic tail risk in real time: Do textual data add value? *International Journal of Forecasting 41*(1), 307–320.

Allen, F. and D. Gale (2000). Financial contagion. *Journal of Political Economy 108*(1), 1–33.

Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). Layer normalization. *arXiv:1607.06450*.

Bai, Y., F. Chen, H. Wang, C. Xiong, and S. Mei (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv:2306.04637*.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *Annals of Statistics 33*(4), 1497–1537.

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics 104*(3), 535–559.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.

Breitung, C. and S. Müller (2025). Global business networks. *Journal of Financial Economics 166*, 104007.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2020, January). The structure of economic news. Working Paper 26648, National Bureau of Economic Research.

Cai, T. T. and R. Ma (2022). Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. *Journal of Machine Learning Research 23*(301), 1–54.

Cao, Y., Z. Chen, P. Kumar, Q. Pei, Y. Yu, H. Li, F. Dimino, L. Ausiello, K. P. Subbalakshmi, and P. M. Ndiaye (2025). Risklabs: Predicting financial risk using large language model based on multimodal and multi-sources data. *arXiv:2404.07452*.

Chao, S.-K., W. K. Härdle, and W. Wang (2015). *Quantile Regression in Risk Calibration*, pp. 1467–1489. New York, NY: Springer New York.

Cordonnier, J.-B., A. Loukas, and M. Jaggi (2020, 01). On the relationship between self-attention and convolution. In *International Conference on Learning Representations*.

Diebold, F. X. and K. Yilmaz (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics 182*(1), 119–134.

Dimitriadis, T. and Y. Hoga (2025). Dynamic CoVaR modeling and estimation. *Journal of Business & Economic Statistics, Forthcoming*.

Ding, X., Y. Zhang, T. Liu, and J. Duan (2014). Using structured events to predict stock price movement: An empirical investigation. In *Conference on Empirical Methods in Natural Language Processing*.

Dougal, C., J. Engelberg, D. García, and C. A. Parsons (2012). Journalists and the stock market. *Review of Financial Studies 25*(3), 639–679.

Edelman, B., S. Goel, S. Kakade, and C. Zhang (2022, June). Inductive biases and variable creation in self-attention mechanisms. In *ICML 2022*.

Engelberg, J. E. and C. A. Parsons (2011). The causal impact of media in financial markets. *Journal of Finance 66*(1), 67–97.

Engle, R. F. (1982, July). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*(4), 987–1007.

Fedyk, A. (2024). Front-page news: The effect of news positioning on financial markets. *Journal of Finance 79*(1), 5–33.

Gai, P., A. Haldane, and S. Kapadia (2011). Complexity, concentration and contagion. *Journal of Monetary Economics 58*(5), 453–470.

Garcia, D. (2013). Sentiment during recessions. *Journal of Finance 68*(3), 1267–1300.

Garg, S., D. Tsipras, P. Liang, and G. Valiant (2022). What can transformers learn in-context? a case study of simple function classes. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Glasserman, P. and C. Lin (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *arXiv:2309.17322*.

Greenwood, R., A. Landier, and D. Thesmar (2015). Vulnerable banks. *Journal of Financial Economics 115*(3), 471–485.

Hautsch, N., J. Schaumburg, and M. Schienle (2014, 03). Financial network systemic risk contributions. *Review of Finance 19*(2), 685–738.

Hayakawa, S. and T. Suzuki (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks 123*, 343–361.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Hull, J. and A. White (1998, September). Incorporating volatility updating into the historical simulation method for VaR. *Journal of Risk 1*(1), 5–19.

Härdle, W. K., W. Wang, and L. Yu (2016). TENET: Tail-event driven network risk. *Journal of Econometrics 192*(2), 499–513. Innovations in Multiple Time Series Analysis.

Jha, M., H. Liu, and A. Manela (2025, 02). Does finance benefit society? A language embedding approach. *Review of Financial Studies*, hhaf012.

Jiao, Y., Y. Lai, D. Sun, Y. Wang, and B. Yan (2025). Approximation bounds for transformer networks with application to regression. *arXiv:2504.12175*.

Ke, Z. T., B. T. Kelly, and D. Xiu (2019, August). Predicting returns with text data. Working Paper 26186, National Bureau of Economic Research.

Keiber, K. L. and H. Samyschew (2017). The world price of sentiment risk. *Global Finance Journal 32*, 62–82.

Keilbar, G. and W. Wang (2022, 01). Modelling systemic risk using neural network quantile regression. *Empirical Economics 62*, 1–26.

Kim, A., M. Muhn, V. Nikolaev, and Y. Zhang (2024). Learning fundamentals from text. *SSRN:10.2139/ssrn.5047947*.

Kim, J., T. Nakamaki, and T. Suzuki (2024). Transformers are minimax optimal nonparametric in-context learners. *arXiv:2408.12186*.

Knight, K. (1998). Limiting distributions for L1 regression estimators under general conditions. *Annals of Statistics 26*(2), 755–770.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Kogan, S., D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, USA, pp. 272–280. Association for Computational Linguistics.

Kusupati, A., G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, and A. Farhadi (2024). Matryoshka representation learning. *arXiv:2205.13147*.

Larsen, V. H. and L. A. Thorsrud (2019). The value of news for economic developments. *Journal of Econometrics 210*(1), 203–218.

Lee, J., F. Chen, S. Dua, D. Cer, M. Shanbhogue, I. Naim, G. H. Ábrego, Z. Li, K. Chen, H. S. Vera, X. Ren, S. Zhang, D. Salz, M. Boratko, J. Han, B. Chen, S. Huang, V. Rao, P. Suganthan, F. Han, A. Doumanoglou, N. Gupta, F. Moiseev, C. Yip, A. Jain, S. Baumgartner, S. Shahi, F. P. Gomez, S. Mariserla, M. Choi, P. Shah, S. Goenka, K. Chen, Y. Xia, K. Chen, S. M. K. Duddu, Y. Chen, T. Walker, W. Zhou, R. Ghiya, Z. Gleicher, K. Gill, Z. Dong, M. Seyedhosseini, Y. Sung, R. Hoffmann, and T. Duerig (2025). Gemini embedding: Generalizable embeddings from Gemini. *arXiv:2503.07891*.

Lee, J., Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, K. Hui, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. H. Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, and I. Naim (2024). Gecko: Versatile text embeddings distilled from large language models. *arXiv:2403.20327*.

Lerner, J. and D. Keltner (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology 81*(1), 146–159.

Lerner, J. S. and D. Keltner (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion 14*(4), 473–493.

Li, H., M. Wang, S. Lu, X. Cui, and P.-Y. Chen (2024). How do nonlinear transformers learn and generalize in in-context learning? *arXiv:2402.15607*.

Li, J., L. Yang, B. Smyth, and R. Dong (2020). MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, New York, NY, USA, pp. 3063–3070. Association for Computing Machinery.

Li, M. Z. F. and R. J. Tomkins (1992). Stability theorems for maxima of dependent sequences. *Canadian Journal of Statistics / La Revue Canadienne de Statistique 20*(2), 211–219.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *Journal of Finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2014). Measuring readability in financial disclosures. *Journal of Finance 69*(4), 1643–1671.

Marek, M., S. Lotfi, A. Somasundaram, A. G. Wilson, and M. Goldblum (2025). Small batch size training for language models: When vanilla SGD works, and why gradient accumulation is wasteful. *arXiv:2507.07101*.

McNeil, A. J. and R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance 7*(3), 271–300. Special issue on Risk Management.

Mousavi-Hosseini, A., C. Sanford, D. Wu, and M. A. Erdogdu (2025). When do Transformers outperform feedforward and recurrent networks? A statistical perspective. *arXiv:2503.11272*.

Nassirtoussi, A. K., S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications 41*(16), 7653–7670.

Omranpour, S., G. Rabusseau, and R. Rabbany (2024). Higher order transformers: Enhancing stock movement prediction on multimodal time-series data. *arXiv:2412.10540*.

Padilla, O. H. M., W. Tansey, and Y. Chen (2022). Quantile regression with ReLU networks: Estimators and minimax rates. *Journal of Machine Learning Research 23*(247), 1–42.

Patton, A. J., J. F. Ziegel, and R. Chen (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics 211*(2), 388–413.

Pele, D. T., V. Bolovăneanu, M.-B. Lin, R. Ren, A. T. Ginavar, B. Spilak, A.-V. Andrei, F.-M. Toma, S. Lessmann, and W. K. Härdle (2026). In the beginning was the word: LLM-VaR and LLM-ES. *Expert Systems with Applications 295*, 128676.

Petrova, G. and P. Wojtaszczyk (2023). Lipschitz widths. *Constructive Approximation 57*, 759–805.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics 48*(4), 1875–1897.

Schuettler, J., F. Audrino, and F. Sigrist (2024, Aug). Does sentiment help in asset pricing? A novel approach using large language models and market-based labels. Swiss Finance Institute Research Paper Series 24-69, Swiss Finance Institute.

Segoviano, M. A. and C. A. E. Goodhart (2009). Banking stability measures. IMF working paper, International Monetary Fund.

Sonkar, S. and R. G. Baraniuk (2023). Investigating the role of feed-forward networks in transformers using parallel attention and feed-forward net design. arXiv preprint arXiv:2305.13297.

Suzuki, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. *Annals of Statistics 47*(5), 2326–2358.

Tetlock, P. (2007, 02). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*, 1139–1168.

Trauger, J. and A. Tewari (2024, 02–04 May). Sequence length independent norm-based generalization bounds for transformers. In S. Dasgupta, S. Mandt, and Y. Li (Eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR.

van der Maaten, L. and G. Hinton (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research 9*(86), 2579–2605.

van Dijk, D. and J. de Winter (2023, March). Nowcasting GDP using tone-adjusted time varying news topics: Evidence from the financial press. Working Papers 766, DNB.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, M. and T. Ma (2024). MANA-Net: Mitigating aggregated sentiment homogenization with news weighting for enhanced market prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2379–2389. Association for Computing Machinery.

Wang, S., M. Cheng, and C. D. Wang (2025). NewsNet-SDF: Stochastic discount factor estimation with pretrained language model news embeddings via adversarial networks. *arXiv:2505.06864*.

Wang, S., T. Ji, L. Wang, Y. Sun, S.-C. Liu, A. Kumar, and C.-T. Lu (2024). Stocktime: A time series specialized large language model architecture for stock price prediction.

Yang, L., T. L. J. Ng, B. Smyth, and R. Dong (2020). HTML: Hierarchical Transformer-based Multi-task Learning for volatility prediction. In *Proceedings of The Web Conference 2020*, WWW '20, New York, NY, USA, pp. 441–451. Association for Computing Machinery.

Yin, W., K. Kann, M. Yu, and H. Schütze (2017). Comparative study of CNN and RNN for natural language processing. *arXiv:1702.01923*.

Zhang, T. (2002, 06). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research 2*, 527–550.

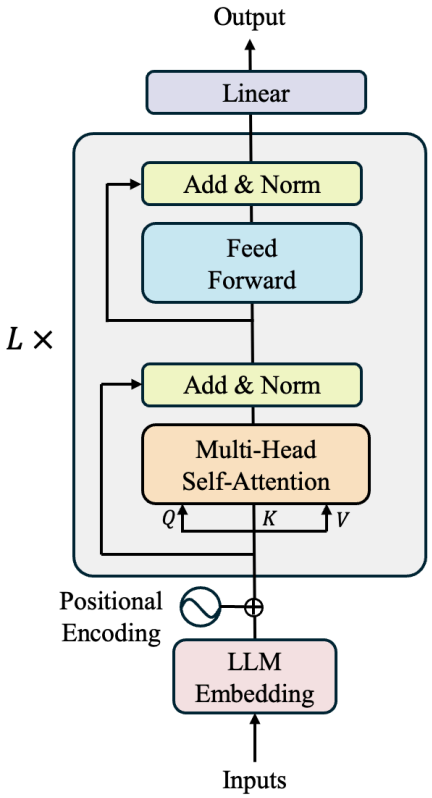
Zhang, Z., K. Ritscher, and O. H. M. Padilla (2025). Quantile additive trend filtering. In Y. Li, S. Mandt, S. Agrawal, and E. Khan (Eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, Volume 258 of *Proceedings of Machine Learning Research*, pp. 4735–4743. PMLR.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen (2025). A survey of large language models. *arXiv:2303.18223*.

Zou, Y. and D. Herremans (2023). PreBit — a multimodal model with twitter FinBERT embeddings for extreme price movement prediction of Bitcoin. *Expert Systems with Applications 233*, 120838.

A Transformer Architect

Figure 12: The Transformer model with residual connection and layer normalization



Notes: The Transformer encoder layer, including the embedding layer, positional encoding, multi-head self-attention (MSA), feed-forward network (FFN), residual connections, layer normalization, and a linear output layer.

A Transformer encoder consists of L layers, each comprising a multi-head self-attention sublayer and a feed-forward network sublayer. It is also possible to add residual connection and layer normalization after each sublayer, which will make the training more stable.

Formally, we define the l-th multi-head self-attention class with residual connection $\widetilde{\mathcal{G}}_l^{(MSA)}(n, d, H)$ as the set of map $\widetilde{g}_l^{(msa)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of the form

$$\widetilde{g}_l^{(msa)}(Z) = \sum_{h=1}^{H} W_{l,h}^{(O)} W_{l,h}^{(V)} Z \cdot \sigma_S \left(\frac{Z^\top {W_{l,h}^{(K)}}^\top W_{l,h}^{(Q)} Z}{\sqrt{d}} \right). \tag{28}$$

For each head $h \in \{1, \ldots, H\}$, we denote by $W_{l,h}^{(V)}, W_{l,h}^{(K)}, W_{l,h}^{(Q)} \in \mathbb{R}^{\frac{d}{H}\times d}$, and $W_{l,h}^{(O)} \in \mathbb{R}^{d\times \frac{d}{H}}$ the value, key, query, and projection matrices, respectively.

The l-th feed-forward network class with residual connection $\widetilde{\mathcal{G}}_l^{(FF)}(d, d_h)$ is defined as the set of maps

$\widetilde{g}_l^{(ff)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of the form

$$\widetilde{g}_l^{(ff)}(Z) = W_l^{(2)}\,\sigma_R\left(W_l^{(1)}\cdot Z + b_l^{(1)}\right) + b_l^{(2)}. \tag{29}$$

where d_h denotes the hidden dimension of the feed-forward network. The parameters are $W_l^{(1)} \in \mathbb{R}^{d_h\times d}$, $b_l^{(1)} \in \mathbb{R}^{d_h}$ and $W_l^{(2)} \in \mathbb{R}^{d\times d_h}$, $b_l^{(2)} \in \mathbb{R}^d$.

In addition, we define the l-th layer normalization function class $\mathcal{G}_l^{(LN)}$ as the set of map $g_l^{(ln)} : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ of the form

$$g_l^{(ln)}(Z) = \left(\frac{Z_j - \frac{1}{d}\sum_i Z_{i,j}}{\sqrt{\frac{1}{d}\sum_k (Z_{k,j} - \frac{1}{d}\sum_i Z_{i,j})}}\right)_{j=1}. \tag{30}$$

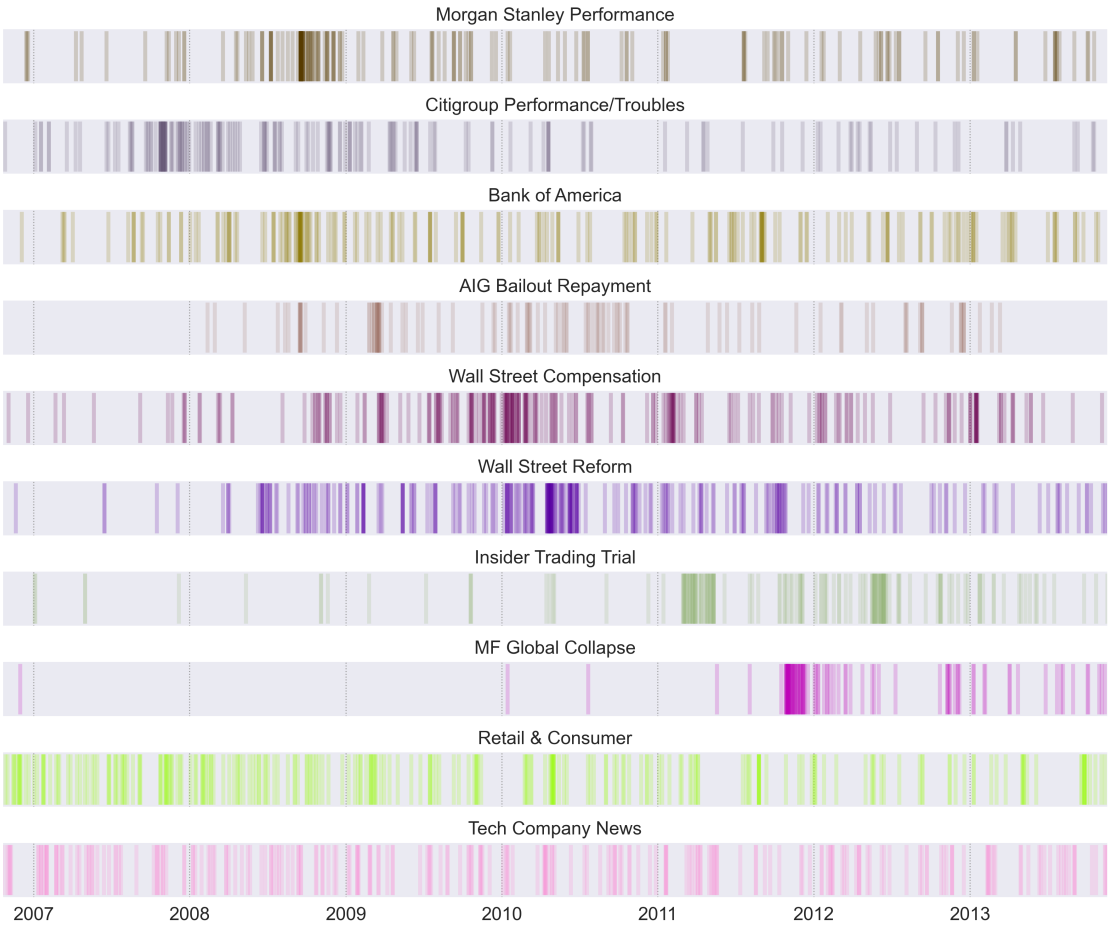
Building on the definition of the Transformer with residual connections and layer normalization, we consider a class of Transformer block of the following form:

$$\begin{aligned}\widetilde{\mathcal{G}}^{(TF)}(n,d,H,d_h,L) := \Big\{ & g_{Lb}^{(ln)} \circ \widetilde{g}_L^{(ff)} \circ g_{La}^{(ln)} \circ \widetilde{g}_L^{(msa)} \circ \cdots \\ & \cdots \circ g_{1b}^{(ln)} \circ \widetilde{g}_1^{(ff)} \circ g_{1a}^{(ln)} \circ \widetilde{g}_1^{(msa)} \Big\}.\end{aligned} \tag{31}$$

Appending an MLP, defined in the same manner as in Section 3.3, yields the Transformer-based neural network class with residual connections and layer normalization.

B News Content

Figure 13: News timeline for selected clusters.



Notes: Each row represents one cluster, with vertical lines indicating the publication dates of news articles belonging to that cluster. Colors correspond to cluster identities, and titles describe the main theme of each cluster. The x-axis shows the timeline, allowing comparison of news activity patterns across topics.

To better understand why the inclusion of news embeddings leads to more negative CoVaR predictions during periods of financial distress, we analyze the underlying content of the news articles used in the model. Our aim is to identify the themes of news and analyze their relationship with the observed changes in systemic risk estimates. While this content analysis can provide insights into narrative-based risk dynamics, we acknowledge the limitation that causality between news and market movements cannot be definitively established. News and financial markets may interact as part of an endogenous system, mutually influencing one another.

Figure 13 illustrates the temporal distribution of major news clusters. Notably, themes such as Morgan Stanley Performance, Citigroup Performance, and Bank of America Troubles dominate the period between

2007 and 2009, with some appearing even earlier. These topics reflect widespread distress among major financial institutions during the global financial crisis. Articles within these clusters frequently use terms such as "losses," "fear," and "woes," highlighting the pessimistic view surrounding the banking sector at the time. The concentration of such narratives likely contributed to the more negative CoVaR predictions during this period.

Another negative dip CoVaR occurs around 2010, coinciding with a surge in news coverage related to Wall Street Compensation and Financial Reform Efforts. In the wake of the crisis, public criticism of executive compensation practices in the financial sector intensified. The Obama administration introduced major financial reform legislation, notably the Dodd-Frank Act, to address perceived regulatory failures. This policy response and public backlash were extensively covered in the media, potentially reinforcing a more negative systemic risk estimates. These clusters underscore the socio-political consequences of the crisis and how they were captured in the news.

In early 2011, coverage of insider trading allegations involving Goldman Sachs appeared to coincide with another reversal in CoVaR trends. Similarly, the MF Global Collapse cluster spanning 2011 to 2012 demonstrates that news about corporate failures continued to influence perceptions of systemic risk well beyond the immediate aftermath of the 2008 crisis. MF Global, a major derivatives broker, filed for bankruptcy following risky bets on European debt, and its collapse was widely reported as a significant market event.

However, not all news clusters exhibit a strong connection to systemic risk. For instance, clusters related to Retail or Technology companies tend to be more evenly distributed throughout the sample period and display relatively weak correlations with the CoVaR estimates of financial institutions. These findings suggest that only certain types of news, especially those focused on financial sector instability, regulation, and corporate misconduct, meaningfully contribute to systemic risk as captured by our embedding-based CoVaR model.

C Monte Carlo Simulation

We validate our empirical Transformer-based model using a Monte Carlo simulation experiment. The goal is to examine the model's ability to learn the dependence between two financial institutions from noisy textual and numerical data in a finite-sample setting. We generate two synthetic time series representing the stock returns of two hypothetical institutions. The return of the second institution is conditional on the return of the first institution.

C.1 Simulation Setting

The baseline setting is linear. The first time series follows a first-order autoregressive process (AR(1)), and the second series is a linear function conditional on the first series with a Gaussian noise:

$$y_t^{(1)} = \phi\, y_{t-1}^{(1)} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_1^2), \tag{32a}$$

$$y_t^{(2)} = \beta\, y_t^{(1)} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_2^2), \tag{32b}$$

with parameters set as $\phi = 0.8$, $\sigma_1 = 0.15$, $\beta = 1.2$, $\sigma_2 = 0.2$, and the initial value $y_{t=0}^{(1)} = 0$. Under this setting, we focus on the CoVaR estimates of the second institution, which is conditional on the first institution being at its VaR level. Notably, we feed the Transformer-based model the real-world *Reuters* financial news as noisy textual input alongside the generated numerical return data. Since these news have no real relation to our synthetic returns, we expect that the Transformer-based model will effectively learn the true numerical dependence, despite the noisy textual input.

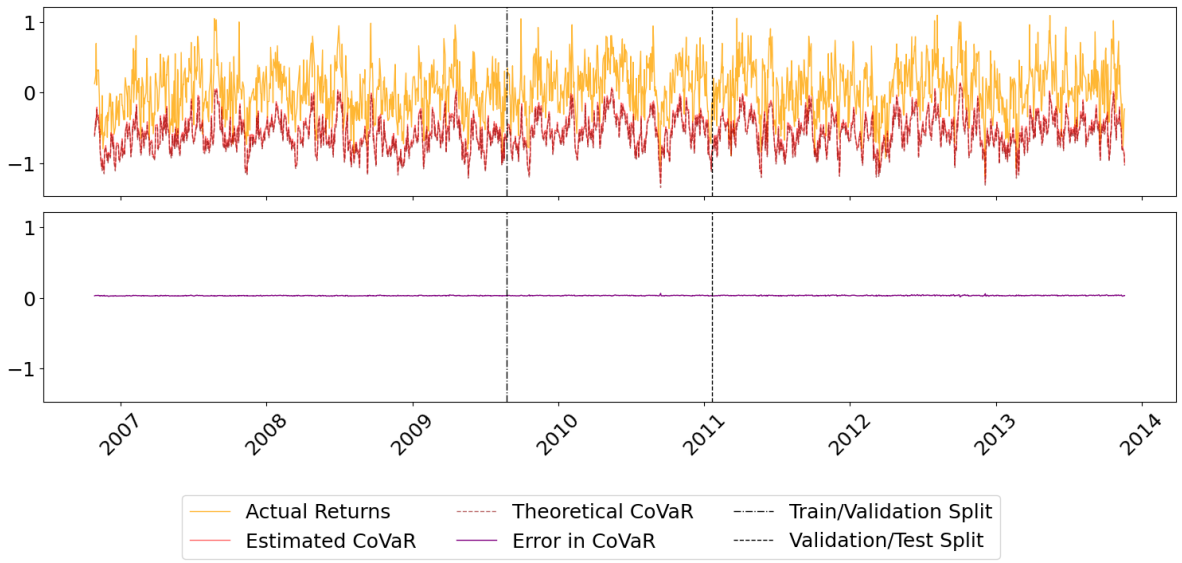
To compute the theoretical CoVaR, we first calculate the VaR for the first institution based on the conditional mean and volatility:

$$\mathrm{VaR}_t^{(1)}(\tau) = \phi\, y_{t-1}^{(1)} + \sigma_1\, z(\tau). \tag{33}$$

where $z(\tau) = \Phi^{-1}(\tau)$ and $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function of the standard normal distribution. We set $\mathrm{VaR}_{t=0}^{(1)}(\tau) = \sigma_1 z(\tau)$ for $t = 0$. Then, we derive the theoretical CoVaR for the second institution by conditioning on the VaR of the first series. Specifically,

$$\mathrm{CoVaR}_t^{(2|1)}(\tau) = \beta\, \mathrm{VaR}_t^{(1)}(\tau) + \sigma_2\, z(\tau). \tag{34}$$

Figure 14: Simulation result



Notes: **Top panel:** Estimated CoVaR from the Transformer-based model. **Bottom panel:** Prediction error of CoVaR. Left and right vertical lines are the train-validation and validation-test splits, respectively.

For comparison purposes, one can also compute the VaR for the second institution:

$$\text{VaR}_t^{(2)}(\tau) = \beta \cdot \phi \, y_{t-1}^{(1)} + \sqrt{\beta^2 \, \sigma_1^2 + \sigma_2^2} \cdot z(\tau), \tag{35}$$

with the initial value set as $\text{VaR}_{t=0}^{(2)} = \sqrt{\beta^2 \, \sigma_1^2 + \sigma_2^2} \cdot z(\tau)$. The VaR of the second institution incorporates randomness from both ϵ_t and η_t.

C.2 Model Predictive Performance

We visualize the simulation results for the Transformer-based model in Figure 14. The estimated CoVaR from this model closely aligns with the theoretical CoVaR, This indicates the model can capture the underlying conditional dependence between the two institutions, even in the presence of extremely noisy textual data. We compute the prediction error, which is the differences between the estimated and true CoVaR. The prediction error for the 5% CoVaR is close to zero.

We benchmark the Transformer-based model against a plain-vanilla MLP consisting of one hidden layer with 64 ReLU-activated units (the model by Keilbar and Wang (2022)). We evaluate the predictive performance of both models using the mean absolute error (MAE) between the estimated CoVaR and its theoretical value

at quantile $\tau = 5\%$:

$$\text{MAE}(\tau) = \frac{1}{T+1}\sum_{t=0}^{T}\big|\widehat{\text{CoVaR}}_t^{(2|1)}(\tau) \;-\; \text{CoVaR}_t^{(2|1)}(\tau)\big|. \tag{36}$$

Table 4 presents the MAE results for different models. For the Transformer-based model, the MAE at the 5% quantiles is close to zero. The MLP models, which rely only on time series inputs, are not affected by the additional noise introduced by the text data. In contrast, the Transformer-based model incorporates noisy textual information, yet it remains capable of capturing the underlying dependency between the two time series. These simulation results support the empirical performance of our Transformer-based model: it does

	5% Quantile
Transformer-based	0.023
MLP (ReLU)	0.039

Table 4: Mean Absolute Error (MAE) at 5% quantile level for the MLP and Transformer.

not merely fit noise but can learn meaningful patterns.

D Proof of Theorem 1

Assumption D.1. *For each $f \in \mathcal{F}$ with the given class $\mathcal{F}$ used for fitting, we have that*

$$\mathbb{E}[f(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^*_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})])^2 \lesssim \mathbb{E}[f(R_{-j,t}, E_{j,t}) - f^*_{j,\tau}(R_{-j,t}, E_{j,t})]^2.$$

Assumption D.1 essentially relates the mean squared error evaluated under VaR values to the ℓ_2 norm bound derived in Lemma F.4, and it holds trivially if the joint density of $(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})$ upon its support is dominated by a constant times the joint density of $(R_{-j,t}, E_{j,t})$.

Lemma D.1. *Suppose that the conclusion of Lemma F.4 applies for the training sample $(X_t, Y_t), t = 1, \cdots, T$ with $X_t = (R_{-j,t}, E_{j,t}), Y_t = R_{j,t}$, $(R_{-j,t'}, E_{j,t'})$ is an independent copy of X_t, $\delta_T > 0, \delta_T \to 0, \delta_T^2 T \to \infty$ as T grows and Assumption D.1 holds then with probability at least $1 - \exp(-\delta_T^2 T)$, then*

$$\mathbb{E}(\widehat{f}_{j,\tau}((VaR^{\tau}_{-j,t'}, E_{j,t'})) - f^*_{j,\tau}((VaR^{\tau}_{-j,t'}, E_{j,t'})))^2 \ \lesssim \ \inf_{f \in \mathcal{F}} \left\| f - f^*_{j,\tau} \right\|^2_{\infty} + \delta_T + \sqrt{\frac{V_{(\mathcal{F}, \|\cdot\|_L \infty)}(\delta_T)}{T}}.$$

Proof. The following is directly implied by the conclusion of Lemma F.4, the definition of $\|\cdot\|_{\ell_2}$ and Assumption D.1: for an arbitrary $r_0 > \delta > 0$, and $\gamma > 0$, with probability at least $1 - \exp(-\gamma)$, we have the following upper bound holds, $\left(\frac{\widetilde{C}_f^2}{c^2} + 1\right) \inf_{f \in \mathcal{F}} \left\| f - f^*_{\tau} \right\|^2_{\infty} + \frac{\widetilde{C}_f}{c} \left(\delta + r_0 \sqrt{\frac{V_{(\mathcal{F}, \|\cdot\|_L \infty)}(\delta)}{T}} + r_0 \sqrt{\frac{\gamma}{T}} + \frac{C_f \gamma}{T} \right)$, where $\widetilde{C}_f, c, V_{(\mathcal{F}, \|\cdot\|_\infty)}(\delta)$ are as specified in Lemma F.4. Then the conclusion follows once we choose $\delta = \delta_T$ and $\gamma = \delta_T^2 T$. □

We now show the proof of Theorem 1.

Proof. The proof proceeds in two parts: approximation error and estimation error.

Recall the definition

$$\widehat{\mathrm{CoVaR}}^{\tau}_{j,t} = \widehat{f}_{j,\theta,\tau}\Big(\widehat{\mathrm{VaR}}^{\tau}_{-j,t},\, E_{j,t}\Big), \qquad \mathrm{CoVaR}^{\tau}_{j,t} = f^*_{j,\tau}\Big(\mathrm{VaR}^{\tau}_{-j,t},\, E_{j,t}\Big). \tag{37}$$

By adding and subtracting $\widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})$ and $f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})$, we obtain

$$\begin{aligned}
&\widehat{f}_{j,\theta,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) \\
&= \Big[\widehat{f}_{j,\theta,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t}, E_{j,t}) - \widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})\Big] \\
&\quad + \Big[\widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})\Big] \\
&\quad + \Big[f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t}, E_{j,t})\Big].
\end{aligned}$$

Taking absolute values and applying the triangle inequality yields

$$\max_{j} \left|\widehat{\mathrm{CoVaR}}^{\tau}_{j,t} - \mathrm{CoVaR}^{\tau}_{j,t}\right| \le A_t + B_t + C_t, \tag{38}$$

where A_t, B_t, C_t denote the absolute values of the three bracketed terms in (38). Under Assumption 2, we have both $f^{*}_{j,\tau}(\cdot, E_{j,t})$ and $\widehat{f}_{j,\theta,\tau}(\cdot, E_{j,t})$ are Lipschitz in their first argument with constants L_r and $\widehat{L}_r$, then

$$\begin{aligned}
A_t &= \max_{j} \left|\widehat{f}_{j,\theta,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t}, E_{j,t}) - \widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})\right| \le \widehat{L}_r \left\|\widehat{\mathrm{VaR}}^{\tau}_{-j,t} - \mathrm{VaR}^{\tau}_{-j,t}\right\|_1, \\
C_t &= \max_{j} \left|f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\widehat{\mathrm{VaR}}^{\tau}_{-j,t}, E_{j,t})\right| \le L_r \left\|\widehat{\mathrm{VaR}}^{\tau}_{-j,t} - \mathrm{VaR}^{\tau}_{-j,t}\right\|_1.
\end{aligned} \tag{39}$$

Assumption 3 implies

$$A_t + C_t = O_p(\widehat{L}_r b_T). \tag{40}$$

Lemma D.1 together with Markov's inequality implies that, for any $M > 0$, with probability at least $1 - \exp(-\delta_T^2 T)$,

$$\mathrm{P}\left(\max_{j} \left|\widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})\right| > M\right) \le \frac{a_T^2}{M^2}, \tag{41}$$

with $a_T^2 = \inf_{f\in\mathcal{F}} \left\|f - f^{*}_{\tau}\right\|^2_{\infty} + \delta_T + \sqrt{\frac{V_{(\mathcal{F},\|\cdot\|_{L^\infty})}(\delta_T)}{T}}$. Hence,

$$B_t = \max_{j} \left|\widehat{f}_{j,\theta,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t}) - f^{*}_{j,\tau}(\mathrm{VaR}^{\tau}_{-j,t}, E_{j,t})\right| = O_p(a_T^2), \tag{42}$$

provided the evaluation is carried out out-of-sample (e.g. via cross-fitting).

Combining terms 1–3 in (38), we obtain

$$\max_{j} \left|\widehat{\mathrm{CoVaR}}^{\tau}_{j,t} - \mathrm{CoVaR}^{\tau}_{j,t}\right| = O_p(a_T^2) + O_p(b_T). \tag{43}$$

By assumption, the VaR error is $O_p(T^{-1/2})$. And the convergence rate of the quantile function error

$\left\|\widehat{f}_{j,N} - f^*_{j,\tau}\right\|_{\ell^2}$ with $\widehat{f}_{j,N}$ the estimator restricted to a finite δ-covering of the function class is shown in Lemma F.4 of the supplementary material. We therefore verify the conditions of Lemma F.4. □

E Proof of Corollary 1

Proof. The proof proceeds by analyzing the trade-off between the approximation bias and the variance associated with the covering number of the architecture. Let $\lceil \cdot \rceil$ denotes the ceiling function and $\vee$ denotes the maximum operator. Define that $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ for two sequences. First, we characterize the Transformer complexity terms η and $\widetilde{\eta}$ in Lemma G.7. From the definitions provided, both terms scale with the cubic root of the logarithmic dimensions of the input:

$$\eta \asymp (\log(dn))^{1/3} \quad \text{and} \quad \widetilde{\eta} \asymp (\log(dn))^{1/3}. \tag{44}$$

Thus, the cubic term appearing in the covering number bound satisfies:

$$(\widetilde{\eta} + \eta)^3 \asymp \log(dn). \tag{45}$$

Based on Theorem 1, the error bound is governed by $\widetilde{R}_{j,\delta,T}$. Thus, by Lemma G.7, H.1, and let $D \asymp m \asymp \log(d_m)$, we obtain

$$\widetilde{R}_{\delta,T} \lesssim_P d_m^{-2\beta/s} + \delta + r_0 \sqrt{\frac{d_m \log(d_m) \cdot \left(\log\left(\delta^{-1} d\right) + \log^2(d_m) \right) + \frac{C_\eta \log(dn)}{\delta^2}}{T}}, \tag{46}$$

where C_η is a constant independent of T, n, and d. Using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have:

$$\widetilde{R}_{\delta,T} \lesssim_P d_m^{-2\beta/s} + \left(\delta + \frac{\sqrt{\log(dn)}}{\delta \sqrt{T}} \right) + \sqrt{\frac{d_m \log(d_m) \cdot \left(\log\left(\delta^{-1} d\right) + \log^2(d_m) \right)}{T}}. \tag{47}$$

We set $\delta \asymp T^{-1/4} (\log(dn))^{1/4}$. Since $D = 8 + (m+5)\left(1 + \lceil \log_2(s \vee \beta) \rceil\right)$, we have $D \asymp m \asymp \log(d_m)$. And d_m will be chosen as a power of T in the balancing step. Then, the bound simplifies to:

$$\widetilde{R}_{\delta,T} \lesssim_P d_m^{-2\beta/s} + T^{-1/4} (\log(dn))^{1/4} + \sqrt{\frac{d_m \log(d_m) \log^3(dT)}{T}}. \tag{48}$$

Pick $d_m \asymp \left(\frac{T}{\log^3(dT)} \right)^{\frac{s}{s+4\beta}}$, and plug this into the error expression, we arrive at the final convergence rate:

$$\max_j \left[\left| \mathrm{CoVaR}_{j,t}^{\tau} - \widehat{\mathrm{CoVaR}}_{j,t}^{\tau} \right| \right] = O_p \left(\left(\frac{T}{\log^3(dT)} \right)^{-\frac{2\beta}{s+4\beta}} \right) + O_p \left(T^{-1/4} (\log(dn))^{1/4} \right) + L_{\mathcal{F}} b_T. \tag{49}$$

This result demonstrates that the rate is governed by the intrinsic dimensionality s and smoothness β.

In addition, the constant $L_{\mathcal{F}}$ characterizes the Lipschitz continuity of the architecture $\mathcal{F}$. Given a one-layer Transformer block with a D-depth MLP (as described in Section H and Theorem 5 in Schmidt-Hieber (2020)), where $D = 8 + (m+5)\left(1 + \lceil \log_2(s \vee \beta) \rceil\right)$, we have

$$L_{\mathcal{F}} \leq \left(L_\sigma^{D-1} \prod_{i=1}^{D} \|W_i\|_\infty \right) \cdot (B_V B_{KQ}). \tag{50}$$

From Lemma H.1, the weights for each head $h \in \{1, \ldots, s\}$ satisfy $\|W_h^{(Q)}\|_{2,1} \leq \frac{\log(4n/\gamma)}{1-2\Delta}$, while other Transformer weights $\|W_h^{(K)}\|_{2,1}, \|W_h^{(V)}\|_{2,1}, \|W_h^{(O)}\|_{2,1} \leq 1$. The MLP weights satisfy $\|W_i\|_\infty \leq B$ for $i = 1, \ldots, D$ and $B > 0$. And $L_\sigma = 1$, we obtain:

$$L_{\mathcal{F}} \leq B^D \left(\frac{s^3 \log(4n/\gamma)}{1-2\Delta} \right) = O(B^D \log(n)). \tag{51}$$

Recall that $D \asymp \log d_m$, and d_m is chosen as a function of T. Looking at the last term in (49), suppose that $b_T = T^{-\alpha}$ with $\alpha \geq \frac{2\beta}{s+4\beta}$. We choose the depth D such that

$$D \asymp \frac{\log d_m}{\log(2 \vee B)} \cdot \frac{\alpha(s+4\beta) - 2\beta}{s}. \tag{52}$$

Then $B^D \leq (2 \vee B)^D \asymp d_m^{(\alpha(s+4\beta)-2\beta)/s}$. And we obtain

$$B^D b_T \lesssim \left(\frac{T}{\log^3(dT)} \right)^{\alpha - \frac{2\beta}{s+4\beta}} T^{-\alpha} = T^{-\frac{2\beta}{s+4\beta}} \cdot (\log(dT))^{-3\left(\alpha - \frac{2\beta}{s+4\beta}\right)}. \tag{53}$$

Therefore, $L_{\mathcal{F}} b_T \lesssim T^{-\frac{2\beta}{s+4\beta}} \log(n)$ up to logarithmic factors. Then the rate in (49) becomes

$$\max_j \left| \mathrm{CoVaR}_{j,t}^{\tau} - \widehat{\mathrm{CoVaR}}_{j,t}^{\tau} \right| = O_p \left(\left(\frac{T}{\log^3(dT)} \right)^{-\frac{2\beta}{s+4\beta}} + T^{-1/4} (\log(dn))^{1/4} + \log(n) \right). \tag{54}$$

□

F Error Bounds for Quantile Regression Estimators

General Setup. In the following proof, let $\{(X_t, Y_t)\}_{t=1}^T$ be an i.i.d. random sample dataset of size T with $(X_t, Y_t) \in \mathbb{R}^{d\times n} \times \mathbb{R}$, and let (X, Y) be independent copy of (X_t, Y_t). The proof is done assuming i.i.d. samples, the extention to, e.g., m-dependent series, are straightforward, see, e.g., Li and Tomkins (1992). Now we consider the target model of the following form:

$$Y_t = f_\tau^*(X_t) + U_{\tau,t}, \tag{55}$$

where $F^{-1}_{U_{\tau,t}|X_t}(\tau) = 0$ and $\tau \in (0, 1)$ is a quantile level. The objective is to approximate $f_\tau^* : \mathbb{R}^{d\times n} \to \mathbb{R}$, given as

$$f_\tau^*(x_t) = F^{-1}_{Y|X=x_t}(\tau), \tag{56}$$

where $F_{Y|X=x_t}$ denotes the conditional distribution of Y given $X = x_t$. We estimate f_τ^* by minimizing the empirical quantile loss over a hypothesis class $\mathcal{F}$, i.e.,

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} \rho_\tau(Y_t - f(X_t)), \tag{57}$$

where $\rho_\tau(u) = u\,(\tau - \mathbf{1}\{u < 0\})$ is the quantile loss function. In addition, let f_T denote the best approximation to f_τ^* within the function class $\mathcal{F}$, defined as

$$f_T := \arg\min_{f\in\mathcal{F}} \mathbb{E}\left[\sum_{t=1}^{T} \rho_\tau\left(Y_t - f(X_t)\right) - \sum_{t=1}^{T} \rho_\tau\left(Y_t - f_\tau^*(X_t)\right)\right]. \tag{58}$$

Next, we state the assumptions and lemmas, followed by the proofs. The proof procedure follows the ideas of Zhang et al. (2025), Padilla et al. (2022).

Assumption F.1 (CDF and Density Boundedness). *There exists a constant $\kappa > 0$ such that for any $\delta = (\delta_1, \cdots, \delta_T) \in \mathbb{R}^T$ satisfying $\|\delta\|_\infty \le \kappa$ we have that*

$$\left|F_{Y|X}(f_\tau^*(X) + \delta_t) - F_{Y|X}(f_\tau^*(X))\right| \ge \underline{p} \cdot |\delta_t|, \quad a.s. \tag{59}$$

for $t = 1, \ldots, T$ and for some constant $\underline{p} > 0$. We also require that

$$\sup_{z\in\mathbb{R}} p_{Y|X}(z) \le \bar{p}, \quad a.s. \tag{60}$$

for some constant $\bar{p} > 0$, where $p_{Y|X}$ is the probability density function of Y conditioning on X.

Lemma F.1. *Suppose Assumption F.1 holds, and assume $\|f\|_\infty \le C_f$ for $f \in \mathcal{F}$. Let $\widetilde{C}_f := \max\{1, 2C_f\}$, then we have*

$$\left\|f - f_T\right\|_{\ell_2}^2 \le \frac{\widetilde{C}_f}{c}\left(\mathbb{E}\Big[\rho_\tau\big(Y - f(X)\big) - \rho_\tau\big(Y - f_T(X)\big)\Big] + \left\|f_T - f_\tau^*\right\|_\infty \left\|f - f_T\right\|_{\ell_2}\right), \tag{61}$$

where $c \gtrsim \min\{\underline{p}, \underline{p}\kappa\}$ is a positive constant, and $\|f_T - f_\tau^*\|_\infty \le c_0$ for a sufficiently small constant c_0.

Proof. Knight (1998)'s identity states that for any two scalars u and w, $\rho_\tau(u - w) - \rho_\tau(u) = -w(\tau - 1\{u \le 0\}) + \int_0^w (1\{u \le s\} - 1\{u \le 0\})\mathrm{d}s$. Let $w_T(X) := f(X) - f_T(X)$. Therefore,

$$\begin{aligned}
\rho_\tau\big(Y - f(X)\big) - \rho_\tau\big(Y - f_T(X)\big) =& -w_T(X)\cdot\Big(\tau - 1\big\{Y \le f_T(X)\big\}\Big) \\
&+ \int_0^{w_T(X)} \Big(1\big\{Y \le f_T(X) + z\big\} - 1\big\{Y \le f_T(X)\big\}\Big)\mathrm{d}z \\
=& -w_T(X)\Big(\tau - 1\big\{Y \le f_\tau^*(X)\big\}\Big) \\
&- w_T(X)\Big(1\big\{Y \le f_\tau^*(X)\big\} - 1\big\{Y \le f_T(X)\big\}\Big) \\
&+ \int_0^{w_T(X)} \Big(1\big\{Y \le f_T(X) + z\big\} - 1\big\{Y \le f_T(X)\big\}\Big)\mathrm{d}z.
\end{aligned} \tag{62}$$

Taking expectations and using Fubini's theorem yields

$$
\begin{aligned}
\mathbb{E}\Big[\rho_\tau\big(Y-f(X)\big)-\rho_\tau\big(Y-f_T(X)\big)\Big] &= \mathbb{E}\Big[-w_T(X)\cdot\Big(\tau-1\big\{Y\le f_\tau^*(X)\big\}\Big)\Big] \\
&\quad -\mathbb{E}\Big[w_T(X)\cdot\Big(1\big\{Y\le f_\tau^*(X)\big\}-1\big\{Y\le f_T(X)\big\}\Big)\Big] \\
&\quad +\mathbb{E}\left[\int_0^{w_T(X)}\Big(1\big\{Y\le f_T(X)+z\big\}-1\big\{Y\le f_T(X)\big\}\Big)\mathrm{d}z\right] \\
&= \mathbb{E}\left[-w_T(X)\cdot\mathbb{E}\Big[\tau-1\big\{Y\le f_\tau^*(X)\big\}\Big|X\Big]\right] \\
&\quad -\mathbb{E}\left[w_T(X)\cdot\mathbb{E}\Big[1\big\{Y\le f_\tau^*(X)\big\}-1\big\{Y\le f_T(X)\big\}\Big|X\Big]\right] \\
&\quad +\mathbb{E}\left[\int_0^{w_T(X)}\mathbb{E}\Big[1\big\{Y\le f_T(X)+z\big\}\Big|X\Big]-\mathbb{E}\Big[1\big\{Y\le f_T(X)\big\}\Big|X\Big]\mathrm{d}z\right] \\
&= \mathbb{E}\left[-w_T(X)\cdot\Big(\tau-F_{Y|X}\big(f_\tau^*(X)\big)\Big)\right] \\
&\quad -\mathbb{E}\left[w_T(X)\cdot\Big(F_{Y|X}(f_\tau^*(X))-F_{Y|X}(f_T(X))\Big)\right] \\
&\quad +\mathbb{E}\left[\int_0^{w_T(X)}\Big(F_{Y|X}(f_T(X)+z)-F_{Y|X}(f_T(X))\Big)\mathrm{d}z\right].
\end{aligned}
\tag{63}
$$

The first term is zero because $f_\tau^*(X)$ is exactly the τ-th quantile. And we can simplify the second term by Jensen's inequality and the triangle inequality:

$$
\begin{aligned}
\left|\mathbb{E}\Big[w_T(X)\cdot\Big(F_{Y|X}(f_\tau^*(X))-F_{Y|X}(f_T(X)\Big)\Big]\right| &\le \mathbb{E}\left[\left|w_T(X)\cdot\Big(F_{Y|X}(f_\tau^*(X))-F_{Y|X}(f_T(X))\Big)\right|\right] \\
&\le \mathbb{E}\Big[\big|w_T(X)\big|\cdot\big|F_{Y|X}(f_\tau^*(X))-F_{Y|X}(f_T(X))\big|\Big] \\
&\le \bar{p}\,\mathbb{E}\Big[\big|w_T(X)\big|\cdot\big|f_\tau^*(X)-f_T(X)\big|\Big],
\end{aligned}
\tag{64}
$$

where $\bar{p}>0$ is the upper bound constant on the conditional probability density from Assumption F.1. For the third term, we need to consider two cases: $|w_T(X)|\le\kappa$ and $|w_T(X)|>\kappa$. When $|w_T(X)|\le\kappa$, the lower bound assumption in Assumption F.1 ensures that the conditional CDF is locally Lipschitz with slope

at least $\underline{p} > 0$. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\int_0^{w_T(X)} \Big(F_{Y|X}(f_T(X)+z) - F_{Y|X}(f_T(X))\Big)\mathrm{d}z\right] &\geq \mathbb{E}\left[\int_0^{w_T(X)} \underline{p}|z|\,\mathrm{d}z\right] \\ &\geq \mathbb{E}\left[\left[\frac{1}{2}\underline{p}z^2\right]_0^{w_T(X)}\right] \\ &= \frac{1}{2}\underline{p}\,\mathbb{E}\Big[w_T^2(X)\Big].\end{aligned} \tag{65}$$

Suppose then that $w_T(X) > \kappa$. Then, we restrict the integration to the interval $[\kappa/2, w_T(X)]$ to ensure that the lower bound applies uniformly. By monotonicity of the CDF, this gives

$$\begin{aligned}&\mathbb{E}\left[\int_0^{w_T(X)} \Big(F_{Y|X}(f_T(X)+z) - F_{Y|X}(f_T(X))\Big)\mathrm{d}z\right] \\ &\geq \mathbb{E}\left[\int_{\kappa/2}^{w_T(X)} \Big(F_{Y|X}(f_T(X)+z) - F_{Y|X}(f_T(X))\Big)\mathrm{d}z\right] \\ &\geq \mathbb{E}\left[(w_T(X)-\kappa/2)\,\Big(F_{Y|X}(f_T(X)+\kappa/2) - F_{Y|X}(f_T(X))\Big)\right] \\ &\geq \mathbb{E}\left[(w_T(X)-\kappa/2)\cdot \underline{p}\,\Big|\frac{\kappa}{2}\Big|\right] \\ &\geq \frac{\underline{p}\kappa}{4}\,\mathbb{E}\big[w_T(X)\big].\end{aligned} \tag{66}$$

The first inequality holds because the integrand is nonnegative for positive z, the second inequality is due to bounding the integral by the product of its length and the minimum value on the interval, the third inequality applies the lower bound from Assumption F.1, and the last inequality uses $w_T(X) > \kappa \Rightarrow w_T(X) - \kappa/2 > w_T(X)/2$. Note that the case $w_T(X) < -\kappa$ is symmetric, and can be covered by taking the absolute value of $w_T(X)$. Combining both regimes, we obtain

$$\begin{aligned}\mathbb{E}\left[\int_0^{w_T(X)} \Big(F_{Y|X}(f_T(X)+z) - F_{Y|X}(f_T(X))\Big)\mathrm{d}z\right] &\geq \mathbb{E}\Big[\min\big\{\frac{\underline{p}}{2}\,w_T^2(X), \frac{\underline{p}\kappa}{4}\,|w_T(X)|\,\big\}\Big] \\ &\geq c\cdot\mathbb{E}\Big[\min\big\{w_T^2(X), |w_T(X)|\,\big\}\Big],\end{aligned} \tag{67}$$

where $c > 0$ and $c \asymp \min\{\underline{p}, \underline{p}\kappa\}$. Recall $w_T(X) = f(X) - f_T(X)$, and $\|w_T(X)\|_\infty \le 2C_f$. Therefore,

$$\begin{aligned}
\min\left\{w_T^2(X), |w_T(X)|\right\} &= w_T^2(X)\cdot \mathbf{1}\{|w_T(X)| \le 1\} + |w_T(X)|\cdot \mathbf{1}\{|w_T(X)| > 1\} \\
&= w_T^2(X)\cdot \mathbf{1}\{|w_T(X)| \le 1\} + \frac{|w_T(X)|}{w_T^2(X)} w_T^2(X)\cdot \mathbf{1}\{|w_T(X)| > 1\} \\
&\ge w_T^2(X)\cdot \mathbf{1}\{|w_T(X)| \le 1\} + \frac{1}{2C_f} w_T^2(X)\cdot \mathbf{1}\{|w_T(X)| > 1\} \\
&\ge \min\{1, \frac{1}{2C_f}\}\cdot w_T^2(X).
\end{aligned} \tag{68}$$

Then, (63) becomes

$$\begin{aligned}
\mathbb{E}\Big[\rho_\tau\big(Y - f(X)\big) - \rho_\tau\big(Y - f_T(X)\big)\Big] &\ge -\,\bar{p}\,\mathbb{E}\Big[|w_T(X)|\cdot|f_\tau^*(X) - f_T(X)|\Big] + \frac{c}{\widetilde{C}_f}\,\mathbb{E}\Big[w_T^2(X)\Big] \\
&\ge -\,\bar{p}\,\sqrt{\mathbb{E}\Big[|w_T(X)|^2\Big]}\sqrt{\mathbb{E}\Big[|f_\tau^*(X) - f_T(X)|^2\Big]} \\
&\quad + \frac{c}{\widetilde{C}_f}\,\mathbb{E}\Big[w_T^2(X)\Big] \\
&\ge -\,\bar{p}\,\Big\|f_T - f_\tau^*\Big\|_\infty \Big\|f - f_T\Big\|_{\ell_2} + \frac{c}{\widetilde{C}_f}\Big\|f - f_T\Big\|_{\ell_2}^2.
\end{aligned} \tag{69}$$

The second inequality is by Cauchy–Schwarz, and $\widetilde{C}_f := \max\{1, 2C_f\}$. Rearranging gives us

$$\Big\|f - f_T\Big\|_{\ell_2}^2 \le \frac{\widetilde{C}_f}{c}\left(\mathbb{E}\Big[\rho_\tau\big(Y - f(X)\big) - \rho_\tau\big(Y - f_T(X)\big)\Big] + \Big\|f_T - f_\tau^*\Big\|_\infty \Big\|f - f_T\Big\|_{\ell_2}\right), \tag{70}$$

where c absorbs the other constants. This completes the proof. □

Lemma F.2. *Suppose Lemma F.1 holds. Let* $M_T(f) := \mathbb{E}\big[\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_T(X))\big]$ *and* $\widehat{M}_T(f)$ *be its empirical counterpart. Then, we have*

$$\Big\|\widehat{f} - f_\tau^*\Big\|_{\ell_2}^2 \le\; 2\left(\frac{\widetilde{C}_f^2}{c^2} + 1\right)\inf_{f\in\mathcal{F}}\Big\|f - f_\tau^*\Big\|_\infty^2 + \frac{4\widetilde{C}_f}{c}\left(M_T(\widehat{f}) - \widehat{M}_T(\widehat{f})\right), \tag{71}$$

where $c \gtrsim \min\{\underline{p}, \underline{p}\kappa\}$, $\widetilde{C}_f = \max\{1, 2C_f\}$, *and* $\inf_{f\in\mathcal{F}}\|f - f_\tau^*\|_\infty^2 < c_0$ *with a sufficiently small* $c_0 > 0$.

Proof. By Lemma F.1 we have that

$$\begin{aligned}\left\|\widehat{f}-f_T\right\|_{\ell_2}^2 &\leq \frac{\widetilde{C}_f}{c}\left(\mathbb{E}\left[\rho_\tau\big(Y-\widehat{f}(X)\big)-\rho_\tau\big(Y-f_T(X)\big)\right]+\left\|f_T-f_\tau^*\right\|_\infty\left\|\widehat{f}-f_T\right\|_{\ell_2}\right)\\ &\leq \frac{\widetilde{C}_f}{c}\left(\mathbb{E}\left[\rho_\tau\big(Y-\widehat{f}(X)\big)-\rho_\tau\big(Y-f_T(X)\big)\right]+\left\|f_T-f_\tau^*\right\|_\infty\left\|\widehat{f}-f_T\right\|_{\ell_2}\right.\\ &\quad\left.-\frac{1}{T}\sum_{t=1}^{T}\rho_\tau\left(y_t-\widehat{f}(x_t)\right)+\frac{1}{T}\sum_{t=1}^{T}\rho_\tau\Big(y_t-f_T(x_t)\Big)\right).\end{aligned} \tag{72}$$

The last inequality follows from the fact $\widehat{f}$ is the global empirical risk minimizer over the sample. Denote

$$M_T(f) := \mathbb{E}\big[\rho_\tau(Y-f(X))-\rho_\tau(Y-f_T(X))\big], \tag{73}$$

and

$$\widehat{M}_T(f) := \frac{1}{T}\sum_{t=1}^{T}\rho_\tau\big(y_t-f(x_t)\big)-\rho_\tau\big(y_t-f_T(x_t)\big). \tag{74}$$

Note that (X,Y) is independent copy of (x_t,y_t). Then, we have

$$\left\|\widehat{f}-f_T\right\|_{\ell_2}^2 \leq \frac{\widetilde{C}_f}{c}\left(M_T(\widehat{f})-\widehat{M}_T(\widehat{f})+\left\|f_T-f_\tau^*\right\|_\infty\left\|\widehat{f}-f_T\right\|_{\ell_2}\right). \tag{75}$$

By Young's inequality,

$$\left\|f_T-f_\tau^*\right\|_\infty\left\|\widehat{f}-f_T\right\|_{\ell_2} \leq \frac{\widetilde{C}_f}{2c}\left\|f_T-f_\tau^*\right\|_\infty^2+\frac{c}{2\widetilde{C}_f}\left\|\widehat{f}-f_T\right\|_{\ell_2}^2. \tag{76}$$

Substituting this bound and rearranging gives

$$\begin{aligned}\left\|\widehat{f}-f_T\right\|_{\ell_2}^2 &\leq \frac{\widetilde{C}_f}{c}\left(M_T(\widehat{f})-\widehat{M}_T(\widehat{f})+\frac{\widetilde{C}_f}{2c}\left\|f_T-f_\tau^*\right\|_\infty^2+\frac{c}{2\widetilde{C}_f}\left\|\widehat{f}-f_T\right\|_{\ell_2}^2\right)\\ &\leq \frac{2\widetilde{C}_f}{c}\Big(M_T(\widehat{f})-\widehat{M}_T(\widehat{f})\Big)+\frac{\widetilde{C}_f^2}{c^2}\left\|f_T-f_\tau^*\right\|_\infty^2.\end{aligned} \tag{77}$$

By the triangle inequality and $(u+v)^2 \leq 2u^2+2v^2$, we obtain

$$\left\|\widehat{f}-f_\tau^*\right\|_{\ell_2}^2 = \left\|\widehat{f}-f_T+f_T-f_\tau^*\right\|_{\ell_2}^2 \leq 2\left\|\widehat{f}-f_T\right\|_{\ell_2}^2+2\left\|f_T-f_\tau^*\right\|_{\ell_2}^2. \tag{78}$$

Replacing $\left\|\widehat{f} - f_T\right\|_{\ell_2}^2$ in (78) with (77) yields

$$\left\|\widehat{f} - f_\tau^*\right\|_{\ell_2}^2 \leq \frac{4\widetilde{C}_f}{c}\left(M_T(\widehat{f}) - \widehat{M}_T(\widehat{f})\right) + 2\left(\frac{\widetilde{C}_f^2}{c^2} + 1\right)\left\|f_T - f_\tau^*\right\|_\infty^2, \tag{79}$$

where we use the fact that the ℓ_2 norm is bounded above by the sup-norm under this case. Finally, since f_T is the best approximation of $f_\tau^* \in \mathcal{F}$, we have $\|f_T - f_\tau^*\|_\infty^2 \leq \inf_{f\in\mathcal{F}} \|f - f_\tau^*\|_\infty^2$. Thus,

$$\left\|\widehat{f} - f_\tau^*\right\|_{\ell_2}^2 \leq 2\left(\frac{\widetilde{C}_f^2}{c^2} + 1\right)\inf_{f\in\mathcal{F}}\left\|f - f_\tau^*\right\|_\infty^2 + \frac{4\widetilde{C}_f}{c}\left(M_T(\widehat{f}) - \widehat{M}_T(\widehat{f})\right). \tag{80}$$

This completes the proof. □

Lemma F.3. *Suppose that Lemmas F.1-F.2 hold. Let $\|\cdot\|_T$ denote the empirical ℓ_2–norm on the sample $\{(X_t, Y_t)\}_{t=1}^T$ of i.i.d. draws from the joint distribution of (X, Y). Suppose $\max\{\|\widehat{f} - f_T\|_{\ell_2},\ \|\widehat{f} - f_T\|_T\} \leq r_0$ for $r_0 > 0$. Then, with probability at least $1 - e^{-\gamma}$, we have*

$$M_T(\widehat{f}) - \widehat{M}_T(\widehat{f}) \lesssim \delta + r_0\sqrt{\frac{V_{(\mathcal{F},\|\cdot\|_\infty)}(\delta)}{T}} + r_0\sqrt{\frac{2\gamma}{T}} + \frac{16C_f\,\gamma}{3T}, \tag{81}$$

for any $\delta \in (0, r_0)$ and any $\gamma > 0$, and $V_{(\mathcal{F},\|\cdot\|_\infty)}(\delta) := \log\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ is the covering number under the sup-norm.

Proof. Denote that $\mathcal{G} := \big\{g : g(x, y) = \rho_\tau\big(y - f(x)\big) - \rho_\tau\big(y - f_T(x)\big),\ f \in \mathcal{F},\ \|f - f_T\|_T \leq r_0\big\}$. Since the quantile loss is 1-Lipschitz[13], we have

$$|g(x, y)| = |\rho_\tau(y - f(x)) - \rho_\tau(y - f_T(x))| \leq |f(x) - f_T(x)| \leq 2C_f. \tag{82}$$

So g maps into the interval $[-2C_f, 2C_f]$. Further, $\left\|\widehat{f} - f_T\right\|_T \leq r_0$ implies that its corresponding $g_{\widehat{f}}$ lies in $\mathcal{G}$, and trivially,

$$M_T(\widehat{f}) - \widehat{M}_T(\widehat{f}) \leq \sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(X, Y)\right] - \frac{1}{T}\sum_{t=1} g(x_t, y_t)\right\}. \tag{83}$$

Then for $\xi_1, \ldots, \xi_T$ independent Rademacher variables independent of $\big\{(x_t, y_t)\big\}_{t=1}^T$, we have the upper

[13]For all real u, v, $|\rho_\tau(u) - \rho_\tau(v)| \leq |u - v|$, because the sub-gradient $\tau - \mathbf{1}\{u < 0\}$ is bounded in $[-1, 1]$.

bound with probability at least $1-\exp(-\gamma)$ by Theorem 2.1 from Bartlett et al. (2005):

$$\sup_{g\in\mathcal{G}}\Big\{\mathbb{E}\left[g(X,Y)\right]-\frac{1}{T}\sum_{t=1}g(x_t,y_t)\Big\}\leq \\ 4\cdot\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\xi_t\,g(x_t,y_t)\;\Big|\;(x_1,y_1),\ldots,(x_T,y_T)\right]+r_0\sqrt{\frac{2\gamma}{T}}+\frac{16C_f\gamma}{3T}. \tag{84}$$

Then, it remains to bound the conditional Rademacher average (i.e., the first term) in the above equation

$$\begin{aligned}\mathbb{E}_\xi\left[\sup_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\xi_t g(x_t,y_t)\right]&\leq\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F},\|f\|_\infty\leq C_f,\|f_T-f\|_T\leq r_0}\frac{1}{T}\sum_{t=1}^{T}\xi_t\big(f(x_t)-f_T(x_t)\big)\right]\\&\leq\inf_{0<\alpha<r_0}\left\{4\alpha+\frac{12}{\sqrt{T}}\int_\alpha^{r_0}\sqrt{\log\mathcal{N}\left(\delta,\mathcal{F},\|\cdot\|_T\right)}\,\mathrm{d}\delta\right\}\\&\leq\inf_{0<\alpha<r_0}\left\{4\alpha+\frac{12}{\sqrt{T}}\int_\alpha^{r_0}\sqrt{\log\mathcal{N}\left(\delta,\mathcal{F},\|\cdot\|_\infty\right)}\,\mathrm{d}\delta\right\}\\&\leq\inf_{0<\alpha<r_0}\left\{4\alpha+\frac{12(r_0-\alpha)}{\sqrt{T}}\sqrt{\log\mathcal{N}\left(\alpha,\mathcal{F},\|\cdot\|_\infty\right)}\right\}\\&\leq\inf_{0<\alpha<r_0}\left\{4\alpha+\frac{12r_0}{\sqrt{T}}\sqrt{\log\mathcal{N}\left(\alpha,\mathcal{F},\|\cdot\|_\infty\right)}\right\}\\&\leq\inf_{0<\alpha<r_0}\left\{4\alpha+12r_0\sqrt{\frac{\log\mathcal{N}\left(\alpha,\mathcal{F},\|\cdot\|_\infty\right)}{T}}\right\}.\end{aligned} \tag{85}$$

The first inequality follows by Talagrand's contraction[14]. We get the second inequality by Dudley's chaining for any $0<\alpha<r_0$, where $\mathcal{N}\left(\delta,\mathcal{F},\|\cdot\|_T\right)$ is the covering number of $\mathcal{F}$ with $\|\cdot\|_T$ balls of radius δ. The third is due to that the empirical norm is smaller than the sup-norm, and the remaining steps use the monotonicity of the covering number in δ and the monotonicity of the integral. Combining the above would lead to the the conclusion. □

Remark F.1. *For Lemma F.3 we only require that the class $\mathcal{G}=\{(x,y)\mapsto\rho_\tau(y-f(x))-\rho_\tau(y-f_T(x)):f\in\mathcal{F}\}$ admits a bounded measurable envelope. Under the uniform bound $\sup_{f\in\mathcal{F}}\|f\|_\infty\leq C_f$ we have $|g|\leq 2C_f$ for all $g\in\mathcal{G}$, so the integrability conditions $\mathbb{E}[\sup_{g\in\mathcal{G}}|g|]<\infty$ and $\mathbb{E}[\sup_{g\in\mathcal{G}}g^2]<\infty$ hold automatically.*

Lemma F.4. *Consider Model (55) with $\tau\in(0,1)$, $u_{\tau,i}$ independent from S_X, then under the notations of Lemmas F.1-F.3, for an arbitrary $r_0>\delta>0$ and $\gamma>0$, with probability at least $1-\exp(-\gamma)$, we have that*

$$\left\|\widehat{f}-f_\tau^*\right\|_{\ell_2}^2\lesssim\left(\frac{\widetilde{C}_f^2}{c^2}+1\right)\inf_{f\in\mathcal{F}}\left\|f-f_\tau^*\right\|_\infty^2+\frac{\widetilde{C}_f}{c}\left(\delta+r_0\sqrt{\frac{V_{(\mathcal{F},\|\cdot\|_{L^\infty})}(\delta)}{T}}+r_0\sqrt{\frac{\gamma}{T}}+\frac{C_f\gamma}{T}\right). \tag{87}$$

[14]Talagrand's contraction: Let $\Phi_1,\cdots,\Phi_T$ be L-Lipschitz functions, $\xi_1,\ldots,\xi_T$ be Rademacher random variables, and $S:=\{x_1,\cdots,x_T\}$ be a random i.i.d. sample. Then, for any hypothesis set $\mathcal{F}$ of real-valued functions, the following inequality holds

$$\frac{1}{T}\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t\cdot\left(\Phi_t\circ f\right)(x_t)\right]\leq\frac{L}{T}\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t\cdot f(x_t)\right]. \tag{86}$$

where $\widetilde{C}_f = \max\{2C_f, 1\}$, $c \asymp \min\{\underline{p}, \underline{p}\kappa\}$, *and* $V_{(\mathcal{F},\|\cdot\|_\infty)}(\delta) := \log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ *is the covering number under the sup-norm.*

Proof. By Lemma F.2, we have

$$\left\|\widehat{f} - f_\tau^*\right\|_{\ell_2}^2 \leq 2\left(\frac{\widetilde{C}_f^2}{c^2} + 1\right) \inf_{f\in\mathcal{F}} \left\|f - f_\tau^*\right\|_\infty^2 + \frac{4\widetilde{C}_f}{c}\left(M_T(\widehat{f}) - \widehat{M}_T(\widehat{f})\right), \tag{88}$$

Plugging the bound of $M_T(\widehat{f}) - \widehat{M}_T(\widehat{f})$ from Lemma F.3 into the above equation, we finish the proof. □

G Constructing the Cover

Let $q \in \mathbb{R}^+$ and consider a function class $\mathcal{F}$, where $f : \mathbb{R}^a \to \mathbb{R}^b$ for $f \in \mathcal{F}$. A subset $\widehat{\mathcal{F}} \subset \mathcal{F}$ is said to cover a collection of inputs $\{x_i\}_{i=1}^n$ if, for every $f \in \mathcal{F}$, there exists some $\widehat{f} \in \widehat{\mathcal{F}}$ such that

$$\sup_{x_i} \left\| f(x_i) - \widehat{f}(x_i) \right\|_q < \epsilon. \tag{89}$$

The smallest size of such a subset is denoted by

$$N_\infty(\mathcal{F}, \epsilon, \{x_i\}_{i=1}^n, \|\cdot\|_q) \tag{90}$$

and is referred to as the empirical covering number with respect to the samples $\{x_i\}_{i=1}^n$. The covering number of $\mathcal{F}$ for n samples is then defined as the worst-case (supremum) over all possible sample sets:

$$N_\infty(\mathcal{F}, \epsilon, n, \|\cdot\|_q) := \sup_{\{x_i\}_{i=1}^n} N_\infty(\mathcal{F}, \epsilon, \{x_i\}_{i=1}^n, \|\cdot\|_q). \tag{91}$$

Following Edelman et al. (2022) and Trauger and Tewari (2024), and recalling the definition and notation of the Transformer in section 3.2, we define the weights of the l-th Transformer layer as

$$W_l := \left\{ W_l^{(KQ)}, W_l^{(V)}, W_l^{(1)}, W_l^{(2)}, b_l^{(1)}, b_l^{(2)} \right\}. \tag{92}$$

Note that since the pairs $\left(W_l^{(K)}, W_l^{(Q)}\right)$ and $\left(W_l^{(O)}, W_l^{(V)}\right)$ are not separately identifiable, we group them together in our notation. And let $W_{1:l} := \{W_1, \ldots, W_{l-1}\}$. For any l, let

$$\begin{aligned} \left\| W_l^{(KQ)} \right\|_2 \le B_{KQ}, \quad \left\| W_l^{(V)} \right\|_2 \le B_V \\ \left\| W_l^{(1)} \right\|_2 \le B_{W1}, \quad \left\| W_l^{(2)} \right\|_2 \le B_{W2}, \end{aligned} \tag{93}$$

and assume $\left\| b_l^{(1)} \right\|_2 \le B_{b1}$ and $\left\| b_l^{(2)} \right\|_2 \le B_{b2}$ for $B_{b1}, B_{b2} > 0$.

We inductively define $g_l^{(tf)} : \mathbb{R}^{d\times n} \mapsto \mathbb{R}^{d\times n}$ starting with $g_1^{(tf)}(X; W_{1:1}) = X$ with $X \in \mathbb{R}^{d\times n}$ the initial input:

$$g_{l+1}^{(tf)}(X; W_{1:l+1}) := \Pi_{norm} \circ g_l^{(ff)} \circ \Pi_{norm} \circ g_l^{(msa)} \circ g_l^{(tf)}\left(X; W_{1:l}\right). \tag{94}$$

where $g_l^{(msa)} \in \mathcal{G}_l^{(MSA)}$ denotes the l-th multi-head self-attention sublayer, $g_l^{(ff)} \in \mathcal{G}_l^{(FF)}$ represents the l-th feed-forward network following the attention sublayer, and Π_{norm} denotes the layer-normalization operator

applied to each column. Adding normalization or not does not affect the final results. We adopt a slightly modified version of the normalization as defined in Section A, following Edelman et al. (2022) and Trauger and Tewari (2024). Specifically, Π_{norm} projects each column of the input onto the unit ball. Unrolling (94) gives us

$$g_{l+1}^{(tf)}\left(X; W_{1:l+1}\right) := \Pi_{norm}\left(\sigma_R\left(\Pi_{norm}\left(g^{(msa)}\left(g_l^{(tf)}\left(X; W_{1:l}\right); W_l\right)\right)W_l^{(1)} + b_l^{(1)}\right)W_l^{(2)} + b_l^{(2)}\right). \tag{95}$$

With the above setups, we derive the following Lemmas.

Lemma G.1. *(Lemma A.8 in Edelman et al. (2022)) For $\epsilon, C_l, \beta_l \geq 0$, and $l \in [1, L]$, the solution to*

$$\min_{\epsilon_1,\ldots,\epsilon_L} \sum_{l=1}^{L} \frac{C_l}{\epsilon_l^2} \quad \text{subject to} \quad \sum_{l=1}^{L} \beta_l \epsilon_l = \epsilon \tag{96}$$

is

$$\frac{\gamma^3}{\epsilon^2}, \tag{97}$$

where

$$\gamma = \sum_{l=1}^{L} C_l^{1/3} \beta_l^{2/3} \quad \text{and} \quad \epsilon_l = \frac{\epsilon}{\gamma}\left(\frac{C_l}{\beta_l}\right)^{1/3}. \tag{98}$$

Proof. The proof is by using Lagrange multipliers. □

Lemma G.2. *Let $\mathcal{B} = \{b \in \mathbb{R}^d : \|b\|_2 \leq B_b\}$ denote the Euclidean ball in $\mathbb{R}^d$ of radius $B_b > 0$. Then, we have*

$$\log N_\infty\left(\varepsilon, \mathcal{B}, \|\cdot\|_2\right) \leq d \log(3B_b) + \frac{d}{\varepsilon^2}, \tag{99}$$

for any $0 < \epsilon \leq B_b$.

Proof. Consider the class of constant functions $\mathcal{F}_b = \{f_b(x) = b : b \in \mathcal{B}\}$. For any $f_b, f_{b'} \in \mathcal{F}_b$, the supremum error over any input set $\mathcal{X}$ is

$$\sup_{x \in \mathcal{X}} \|f_b(x) - f_{b'}(x)\|_2 = \sup_{x \in \mathcal{X}} \|b - b'\|_2 = \|b - b'\|_2. \tag{100}$$

Hence, covering the function class $\mathcal{F}_b$ under the supremum norm is equivalent to covering the Euclidean ball $\mathcal{B}$. The unit ball in $\mathbb{R}^d$ can be covered by at most $(3/\varepsilon)^d$ balls of radius ε. Scaling by B_b gives

$$N_\infty(\varepsilon, \mathcal{B}, \|\cdot\|_2) \leq \left(\frac{3B_b}{\varepsilon}\right)^d. \tag{101}$$

Taking logarithms yields

$$\log N_\infty\big(\varepsilon, \mathcal{B}, \|\cdot\|_2\big) \le d \log\left(\frac{3B_b}{\varepsilon}\right) = d\log(3B_b) + d\log\frac{1}{\varepsilon}. \tag{102}$$

Lastly, note that for any $a > 0$, $\log(1/\varepsilon) = o(1/\varepsilon^a)$ as $\varepsilon \to 0^+$. Choosing $a = 2$, we obtain the looser but convenient bound

$$\log N_\infty\big(\varepsilon, \mathcal{B}, \|\cdot\|_2\big) \le d\log(3B_b) + \frac{d}{\varepsilon^2}. \tag{103}$$

□

Lemma G.3 (Rework of Lemma A.15 in Edelman et al. (2022))**.** *Suppose $W_{1:l+1}, \widehat{W}_{1:l+1}$ satisfy the norm bounds $\left\|W_l^{(KQ)}\right\|_2 \le B_{KQ}$, $\left\|W_l^{(V)}\right\|_2 \le B_V$, $\left\|W_l^{(1)}\right\|_2 \le B_{W1}$, $\left\|W_l^{(2)}\right\|_2 \le B_{W2}$ for all l. Then we have*

$$\begin{aligned}
&\left\|g_{l+1}^{(tf)}\left(X; W_{1:l+1}\right) - g_{l+1}^{(tf)}\left(X; \widehat{W}_{1:l+1}\right)\right\|_{2,\infty} \\
\le &\left\|\left(\sigma_R\left(\Pi_{norm}\left(g^{(msa)}\left(g_l^{(tf)}\left(X; \widehat{W}_{1:l}\right); W_l\right)\right) W_l^{(1)} + b_l^{(1)}\right)\right)\left(W_l^{(2)} - \widehat{W}_l^{(2)}\right)\right\|_{2,\infty} \\
&+ \left\|b_l^{(2)} - \widehat{b}_l^{(2)}\right\|_{2,\infty} \\
&+ B_{W2}L_\sigma\left\|\Pi_{norm}\left(g^{(msa)}\left(g_l^{(tf)}\left(X; \widehat{W}_{1:l}\right); \widehat{W}_l\right)\right)\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} \\
&+ B_{W2}L_\sigma\left\|b_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty} \\
&+ B_{W2}L_\sigma B_{W1}B_V\left(1 + 4B_{KQ}\right)\left\|g_l^{(tf)}\left(X; W_{1:l}\right) - g_l^{(tf)}\left(X; \widehat{W}_{1:l}\right)\right\|_{2,\infty} \\
&+ 2B_{W2}L_\sigma B_{W1}B_V\left\|\left(W^{(KQ)} - \widehat{W}^{(KQ)}\right)^\top g_l^{(tf)}\left(X; \widehat{W}_{1:l}\right)\right\|_{2,\infty} \\
&+ B_{W2}L_\sigma B_{W1}\left\|\left(W^{(V)} - \widehat{W}^{(V)}\right) g_l^{(tf)}\left(X; \widehat{W}_{1:l}\right)\right\|_{2,\infty}.
\end{aligned} \tag{104}$$

Proof. Let

$$g_l := g_l^{(tf)}\left(X; W_{1:l}\right), \quad A_l := \sigma_R\left(\Pi_{norm}\left(g^{(msa)}\left(g_l; W_l\right)\right) W_l^{(1)} + b_l^{(1)}\right). \tag{105}$$

By Lemma A.9 in Edelman et al. (2022), we have

$$
\begin{aligned}
&\left\|g_{l+1}^{(tf)}\left(X; W_{1:l+1}\right) - g_{l+1}^{(tf)}\left(X; \widehat{W}_{1:l+1}\right)\right\|_{2,\infty} \\
=& \left\|\Pi_{norm}\left(A_l W_l^{(2)} + b_l^{(2)}\right) - \Pi_{norm}\left(\widehat{A}_l \widehat{W}_l^{(2)} + \widehat{b}_l^{(2)}\right)\right\|_{2,\infty} \\
\leq& \left\|A_l W_l^{(2)} + b_l^{(2)} - \widehat{A}_l \widehat{W}_l^{(2)} - \widehat{b}_l^{(2)}\right\|_{2,\infty} \\
=& \left\|A_l W_l^{(2)} - \widehat{A}_l \widehat{W}_l^{(2)} + \widehat{A}_l W_l^{(2)} - \widehat{A}_l W_l^{(2)} + b_l^{(2)} - \widehat{b}_l^{(2)}\right\|_{2,\infty} \\
\leq& \left\|\widehat{A}_l\left(W_l^{(2)} - \widehat{W}_l^{(2)}\right)\right\|_{2,\infty} + \left\|\left(A_l - \widehat{A}_l\right) W_l^{(2)}\right\|_{2,\infty} + \left\|b_l^{(2)} - \widehat{b}_l^{(2)}\right\|_{2,\infty},
\end{aligned}
\tag{106}
$$

where the last inequality follows from the triangle inequality. We take a closer look at the second term. By the spectral norm $\left\|W_l^{(2)}\right\|_2 \leq B_{W2}$, we have

$$
\left\|\left(A_l - \widehat{A}_l\right) W_l^{(2)}\right\|_{2,\infty} \leq \left\|W_l^{(2)}\right\|_2 \left\|A_l - \widehat{A}_l\right\|_{2,\infty} \leq B_{W2} \left\|A_l - \widehat{A}_l\right\|_{2,\infty}. \tag{107}
$$

Then we need to bound $\left\|A_l - \widehat{A}_l\right\|_{2,\infty}$. By Lemma A.9 in Edelman et al. (2022) and the assumption that σ_R is L_σ-Lipschitz, we have

$$
\begin{aligned}
&\left\|A_l - \widehat{A}_l\right\|_{2,\infty} \\
\leq& L_\sigma \left\|\Pi_{norm}\left(g^{(msa)}\left(g_l; W_l\right)\right) W_l^{(1)} + b_l^{(1)} - \Pi_{norm}\left(g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right) \widehat{W}_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty} \\
\leq& L_\sigma \left\|\Pi_{norm}\left(g^{(msa)}\left(g_l; W_l\right)\right) W_l^{(1)} - \Pi_{norm}\left(g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right) \widehat{W}_l^{(1)}\right\|_{2,\infty} \\
&+ L_\sigma \left\|b_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty}
\end{aligned}
\tag{108}
$$

Looking at the first term, let

$$
B_l := \Pi_{norm}\left(g^{(msa)}\left(g_l; W_l\right)\right), \tag{109}
$$

then

$$
\begin{aligned}
&L_\sigma \left\|B_l W_l^{(1)} - \widehat{B}_l \widehat{W}_l^{(1)}\right\|_{2,\infty} \\
\leq& L_\sigma \left\|\widehat{B}_l\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} + L_\sigma \left\|\left(B_l - \widehat{B}_l\right) W_l^{(1)}\right\|_{2,\infty} \\
\leq& L_\sigma \left\|\widehat{B}_l\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} + L_\sigma B_{W1} \left\|B_l - \widehat{B}_l\right\|_{2,\infty}.
\end{aligned}
\tag{110}
$$

The last inequality uses the same idea as above by the spectral norm $\left\|W_l^{(1)}\right\|_2 \leq B_{W1}$. Next, we bound $\left\|B_l - \widehat{B}_l\right\|_{2,\infty}$ by Lemma A.9 in Edelman et al. (2022) as the following

$$\begin{aligned}
&\left\|B_l - \widehat{B}_l\right\|_{2,\infty} \\
\leq &\left\|g^{(msa)}\left(g_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right\|_{2,\infty} \\
\leq &\left\|g^{(msa)}\left(g_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; W_l\right) + g^{(msa)}\left(\widehat{g}_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right\|_{2,\infty} \\
\leq &\left\|g^{(msa)}\left(g_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; W_l\right)\right\|_{2,\infty} + \left\|g^{(msa)}\left(\widehat{g}_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right\|_{2,\infty}
\end{aligned} \tag{111}$$

where the third inequality follows from the triangle inequality. For the first term in (111), applying Lemma A.14 in Edelman et al. (2022), we have, for $\left\|W_l^{(V)}\right\|_2 \leq B_V$, $\left\|W_l^{(KQ)}\right\|_2 \leq B_{KQ}$,

$$\left\|g^{(msa)}\left(g_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; W_l\right)\right\|_{2,\infty} \leq B_V\left(1 + 4B_{KQ}\right)\left\|g_l - \widehat{g}_l\right\|_{2,\infty}. \tag{112}$$

For the second term in (111), by Lemma A.13 in Edelman et al. (2022), we have

$$\begin{aligned}
&\left\|g^{(msa)}\left(\widehat{g}_l; W_l\right) - g^{(msa)}\left(\widehat{g}_l; \widehat{W}_l\right)\right\|_{2,\infty} \\
\leq &\, 2B_V\left\|\left(W^{(KQ)} - \widehat{W}^{(KQ)}\right)^\top \widehat{g}_l\right\|_{2,\infty} + \left\|\left(W^{(V)} - \widehat{W}^{(V)}\right)\widehat{g}_l\right\|_{2,\infty}.
\end{aligned} \tag{113}$$

Thus, (111) becomes

$$\begin{aligned}
\left\|B_l - \widehat{B}_l\right\|_{2,\infty} \leq &\, B_V\left(1 + 4B_{KQ}\right)\left\|g_l - \widehat{g}_l\right\|_{2,\infty} \\
&+ 2B_V\left\|\left(W^{(KQ)} - \widehat{W}^{(KQ)}\right)^\top \widehat{g}_l\right\|_{2,\infty} + \left\|\left(W^{(V)} - \widehat{W}^{(V)}\right)\widehat{g}_l\right\|_{2,\infty}.
\end{aligned} \tag{114}$$

Then (108) becomes

$$
\begin{aligned}
&\left\|A_l - \widehat{A}_l\right\|_{2,\infty} \\
&\le L_\sigma \left\|\widehat{B}_l\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} + L_\sigma B_{W1}\left\|B_l - \widehat{B}_l\right\|_{2,\infty} + L_\sigma \left\|b_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty} \\
&\le L_\sigma \left\|\widehat{B}_l\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} + L_\sigma \left\|b_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty} \\
&\quad + L_\sigma B_{W1} B_V \left(1 + 4B_{KQ}\right)\left\|g_l - \widehat{g}_l\right\|_{2,\infty} \\
&\quad + 2L_\sigma B_{W1} B_V \left\|\left(W^{(KQ)} - \widehat{W}^{(KQ)}\right)^\top \widehat{g}_l\right\|_{2,\infty} \\
&\quad + L_\sigma B_{W1}\left\|\left(W^{(V)} - \widehat{W}^{(V)}\right)\widehat{g}_l\right\|_{2,\infty}
\end{aligned}
\tag{115}
$$

Combining the above and plugging into (106) gives us

$$
\begin{aligned}
&\left\|g_{l+1}^{(tf)}\left(X; W_{1:l+1}\right) - g_{l+1}^{(tf)}\left(X; \widehat{W}_{1:l+1}\right)\right\|_{2,\infty} \\
\le\ &\left\|\widehat{A}_l\left(W_l^{(2)} - \widehat{W}_l^{(2)}\right)\right\|_{2,\infty} + \left\|b_l^{(2)} - \widehat{b}_l^{(2)}\right\|_{2,\infty} \\
&+ B_{W2} L_\sigma \left\|\widehat{B}_l\left(W_l^{(1)} - \widehat{W}_l^{(1)}\right)\right\|_{2,\infty} \\
&+ B_{W2} L_\sigma \left\|b_l^{(1)} - \widehat{b}_l^{(1)}\right\|_{2,\infty} \\
&+ B_{W2} L_\sigma B_{W1} B_V \left(1 + 4B_{KQ}\right)\left\|g_l - \widehat{g}_l\right\|_{2,\infty} \\
&+ 2B_{W2} L_\sigma B_{W1} B_V \left\|\left(W^{(KQ)} - \widehat{W}^{(KQ)}\right)^\top \widehat{g}_l\right\|_{2,\infty} \\
&+ B_{W2} L_\sigma B_{W1}\left\|\left(W^{(V)} - \widehat{W}^{(V)}\right)\widehat{g}_l\right\|_{2,\infty}.
\end{aligned}
\tag{116}
$$

□

Lemma G.4 (Rework of Lemma A.16 in Edelman et al. (2022))**.** *Suppose $W_{1:L+1}, \widehat{W}_{1:L+1}$ satisfy our norm bounds. Then we have*

$$
\begin{aligned}
&\left|g^{(tfw)}\left(X; W_{1:L+1}, w\right) - g^{(tfw)}\left(X; \widehat{W}_{1:L+1}, \widehat{w}\right)\right| \\
\le\ &\|w\|_2 \cdot \left\|g_{L+1}^{(tf)}\left(X; W_{1:L+1}\right) - g_{L+1}^{(tf)}\left(X; \widehat{W}_{1:L+1}\right)\right\|_2 \\
&+ \left|\left(g_{L+1}^{(tf)}\left(X; \widehat{W}_{1:L+1}\right) - g_{L+1}^{(tf)}\left(X; \widehat{W}_{1:L+1}\right)\right) \cdot \left(w - \widehat{w}\right)\right|.
\end{aligned}
\tag{117}
$$

Proof. In the last layer, we have $W_{L+1}^{(2)} \in \mathbb{R}^{d \times d_h}, b_{L+1}^{(2)} \in \mathbb{R}^d$. We add a linear mapping with the weight

$w \in \mathbb{R}^n$,

$$
\begin{aligned}
&\left|g_{L+1}^{(tfw)}\big(X; W_{1:L+1}, w\big) - g_{L+1}^{(tfw)}\big(X; \widehat{W}_{1:L+1}, \widehat{w}\big)\right| \\
=\ &\left|g_{L+1}^{(tf)}\big(X; W_{1:L+1}\big)\cdot w - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\widehat{w}\right| \\
=\ &\left|g_{L+1}^{(tf)}\big(X; W_{1:L+1}\big)\cdot w - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot w + g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot w - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot \widehat{w}\right| \\
\leq\ &\left|\Big(g_{L+1}^{(tf)}\big(X; W_{1:L+1}\big) - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\Big)\cdot w\right| + \left|g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot\Big(w - \widehat{w}\Big)\right| \\
\leq\ &\|w\|_2 \cdot \left\|g_{L+1}^{(tf)}\big(X; W_{1:L+1}\big) - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\right\|_2 + \left|g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot\Big(w - \widehat{w}\Big)\right|.
\end{aligned}
\tag{118}
$$

The first inequality uses the triangle inequality. The second inequality uses the inner product by norms. □

Lemma G.5 (Rework of Theorem 4.2 in Trauger and Tewari (2024))**.** *Suppose we have a log covering numbers in the form of C_1/ϵ^2 and C_{B_x}/ϵ^2 for the function class $\{x \mapsto Wx \mid x \in \mathcal{X}, W \in \mathcal{W}\}$ where $\|x\|_2 \leq 1, \forall x \in \mathcal{X}$ and $\|x\|_2 \leq B_x, \forall x \in \mathcal{X}$ respectively. Suppose we also have $\left\|W_l^{(2)}\right\|_2 \leq B_{W2}$, $\left\|W_l^{(1)}\right\|_2 \leq B_{W1}$, $\left\|W_l^{(V)}\right\|_2 \leq B_V$, $\left\|W_l^{(QK)}\right\|_2 \leq B_{QK}$, $\|w\|_2 \leq B_w$, $\left\|b_l^{(1)}\right\|_1 \leq B_{b1}$, $\left\|b_l^{(2)}\right\|_1 \leq B_{b2}$. Let $\alpha_l := \prod_{j=l+1}^{L} B_{W2} L_\sigma B_{W1} B_V \left(1 + 4B_{KQ}\right)$ and*

$$
\begin{aligned}
\beta_l^{(2)} &:= B_w \alpha_l, \quad \beta_l^{(1)} := B_w \alpha_l \cdot B_{W2} L_\sigma, \\
\beta_l^{(KQ)} &:= B_w \alpha_l \cdot 2B_{W2} L_\sigma B_{W1} B_V, \quad \beta_l^{(V)} := B_w \alpha_l \cdot B_{W2} L_\sigma B_{W1}.
\end{aligned}
\tag{119}
$$

Then the covering number of $g_{L+1}^{(tfw)}$ is

$$
\frac{(\widetilde{\eta} + \eta)^3}{\epsilon^2}. \tag{120}
$$

where

$$
\begin{aligned}
\eta &= C_1^{\frac{1}{3}} \sum_{l=2}^{L} \Big(2\beta_l^{(2)^{\frac{2}{3}}} + 2\beta_l^{(1)^{\frac{2}{3}}} + \beta_l^{(KQ)^{\frac{2}{3}}} + \beta_l^{(V)^{\frac{2}{3}}}\Big), \\
\widetilde{\eta} &= C_{B_x}^{\frac{1}{3}} \Big(\beta_1^{(KQ)}\Big) + C_1^{\frac{1}{3}} \Big(2\beta_1^{(2)^{\frac{2}{3}}} + 2\beta_1^{(1)^{\frac{2}{3}}} + 2\beta_1^{(V)^{\frac{2}{3}}}\Big) + C_1^{\frac{1}{3}}.
\end{aligned}
\tag{121}
$$

Proof. By Lemma G.4, we know

$$
\begin{aligned}
&\left|g^{(tfw)}\big(X; W_{1:L+1}, w\big) - g^{(tfw)}\big(X; \widehat{W}_{1:L+1}, \widehat{w}\big)\right| \\
\leq\ &\|w\|_2 \cdot \left\|g_{L+1}^{(tf)}\big(X; W_{1:L+1}\big) - g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\right\|_2 + \left|g_{L+1}^{(tf)}\big(X; \widehat{W}_{1:L+1}\big)\cdot\Big(w - \widehat{w}\Big)\right|.
\end{aligned}
\tag{122}
$$

Thus, we have

$$\begin{aligned}&\left|g^{(tfw)}\left(X;W_{1:L+1},w\right)-g^{(tfw)}\left(X;\widehat{W}_{1:L+1},\widehat{w}\right)\right|\\&\leq\|w\|_2\cdot\left\|g_{L+1}^{(tf)}\left(X;W_{1:L+1}\right)-g_{L+1}^{(tf)}\left(X;\widehat{W}_{1:L+1}\right)\right\|_2+\epsilon^{(w)}.\end{aligned}\tag{123}$$

Next, by Lemma G.3, we have

$$\begin{aligned}&\left\|g_{l+1}^{(tf)}\left(X;W_{1:l+1}\right)-g_{l+1}^{(tf)}\left(X;\widehat{W}_{1:l+1}\right)\right\|_{2,\infty}\\&\leq\left\|\left(\sigma_R\left(\Pi_{norm}\left(g^{(msa)}\left(g_l^{(tf)}\left(X;\widehat{W}_{1:l}\right);W_l\right)\right)W_l^{(1)}+b_l^{(1)}\right)\right)\left(W_l^{(2)}-\widehat{W}_l^{(2)}\right)\right\|_{2,\infty}\\&\quad+\left\|b_l^{(2)}-\widehat{b}_l^{(2)}\right\|_{2,\infty}\\&\quad+B_{W2}L_\sigma\left\|\Pi_{norm}\left(g^{(msa)}\left(g_l^{(tf)}\left(X;\widehat{W}_{1:l}\right);\widehat{W}_l\right)\right)\left(W_l^{(1)}-\widehat{W}_l^{(1)}\right)\right\|_{2,\infty}\\&\quad+B_{W2}L_\sigma\left\|b_l^{(1)}-\widehat{b}_l^{(1)}\right\|_{2,\infty}\\&\quad+B_{W2}L_\sigma B_{W1}B_V\left(1+4B_{KQ}\right)\left\|g_l^{(tf)}\left(X;W_{1:l}\right)-g_l^{(tf)}\left(X;\widehat{W}_{1:l}\right)\right\|_{2,\infty}\\&\quad+2B_{W2}L_\sigma B_{W1}B_V\left\|\left(W^{(KQ)}-\widehat{W}^{(KQ)}\right)^\top g_l^{(tf)}\left(X;\widehat{W}_{1:l}\right)\right\|_{2,\infty}\\&\quad+B_{W2}L_\sigma B_{W1}\left\|\left(W^{(V)}-\widehat{W}^{(V)}\right)g_l^{(tf)}\left(X;\widehat{W}_{1:l}\right)\right\|_{2,\infty}.\end{aligned}\tag{124}$$

Notice that if $X_t\in\mathbb{R}^{d\times n}$ with $t\in[1,T]$, and $W\in\mathbb{R}^{d\times d}$

$$\max_{t\in[1,T]}\left\|(W-\widehat{W})X_t\right\|_{2,\infty}=\max_{t\in[1,T],i\in[1,n]}\left\|(W-\widehat{W})X_{t,i}\right\|\tag{125}$$

Therefore we can use the covering number bounds to bound the values in the $\|\cdot\|_{2,\infty}$.

First, we create a cover for multilayer Transformer with inputs $X_1,\ldots,X_T$ and $\|X_t\|_2\leq B_x$. Let $\mathcal{W}_l^{(V)}$, $\mathcal{W}_l^{(KQ)}$, $\mathcal{W}_l^{(1)}$, $\mathcal{W}_l^{(2)}$, $\mathcal{B}_l^{(1)}$, $\mathcal{B}_l^{(2)}$ be the sets of all possible values for $W_l^{(V)}$, $W_l^{(KQ)}$, $W_l^{(1)}$, $W_l^{(2)}$, $b_l^{(1)}$, $b_l^{(2)}$ respectively. Let $\widehat{\mathcal{W}}_l^{(V)}$ cover the function class

$$\left\{x\mapsto {W^{(V)}}^\top x\mid W^{(V)}\in\mathcal{W}_l^{(V)},\|x\|_2\leq 1\right\}.\tag{126}$$

Let $\widehat{\mathcal{W}}_l^{(KQ)}$ cover the function class

$$\left\{x\mapsto W^{(KQ)}x\mid W^{(KQ)}\in\mathcal{W}_l^{(KQ)},\|x\|_2\leq 1\right\}\tag{127}$$

except for $\widehat{\mathcal{W}}_1^{(KQ)}$, which covers the same function class, but with $\|x\| \leq B_x$. Let $\widehat{\mathcal{W}}_l^{(1)}$ cover the function class

$$\left\{x \mapsto {W^{(1)}}^\top x \mid W^{(1)} \in \mathcal{W}_l^{(1)}, \|x\|_2 \leq 1\right\}. \tag{128}$$

And let $\widehat{\mathcal{W}}_l^{(2)}$ cover the function class

$$\left\{x \mapsto {W^{(2)}}^\top x \mid W^{(2)} \in \mathcal{W}_l^{(2)}, \|x\|_2 \leq 1\right\}. \tag{129}$$

Let $\widehat{\mathcal{B}}_l^{(1)}, \widehat{\mathcal{B}}_l^{(2)}$ cover $b_l^{(1)}, b_l^{(2)}$, respectively. Let all of these classes be covered with Tn points and let $\epsilon_l^{(V)}, \epsilon_l^{(KQ)}, \epsilon_l^{(W1)}, \epsilon_l^{(W2)}, \epsilon_l^{(b1)}, \epsilon_l^{(b2)}$ be the resolution for each cover. Also, let $\widehat{\mathcal{W}}$ cover

$$\left\{x \mapsto w^\top x \mid w \in \mathcal{W}, \|x\|_2 \leq 1\right\} \tag{130}$$

at resolution $\epsilon^{(w)}$. Thus, our final cover is

$$\begin{aligned}\widehat{W}^{1:L+1} \in \widehat{\mathcal{W}}_1^{(KQ)} \otimes \widehat{\mathcal{W}}_1^{(V)} \otimes \widehat{\mathcal{W}}_1^{(1)} \otimes \widehat{\mathcal{B}}_1^{(1)} \otimes \widehat{\mathcal{W}}_1^{(2)} \otimes \widehat{\mathcal{B}}_1^{(2)} \otimes \cdots \\ \cdots \otimes \widehat{\mathcal{W}}_L^{(KQ)} \otimes \widehat{\mathcal{W}}_L^{(V)} \otimes \widehat{\mathcal{W}}_L^{(1)} \otimes \widehat{\mathcal{B}}_L^{(1)} \otimes \widehat{\mathcal{W}}_L^{(2)} \otimes \widehat{\mathcal{B}}_L^{(2)} \otimes \widehat{\mathcal{W}}.\end{aligned} \tag{131}$$

Then (124) becomes

$$\begin{aligned}&\left\|g_{L+1}^{(tf)}\left(X; W_{1:L+1}\right) - g_{L+1}^{(tf)}\left(X; \widehat{W}_{1:L+1}\right)\right\|_{2,\infty} \\ &\leq \epsilon_L^{(W2)} + \epsilon_L^{(b2)} + B_{W2} L_\sigma \epsilon_L^{(W1)} + B_{W2} L_\sigma \epsilon_L^{(b1)} \\ &\quad + B_{W2} L_\sigma B_{W1} B_V \left(1 + 4B_{KQ}\right) \left\|g_L^{(tf)}\left(X; W_{1:L}\right) - g_L^{(tf)}\left(X; \widehat{W}_{1:L}\right)\right\|_{2,\infty} \\ &\quad + 2B_{W2} L_\sigma B_{W1} B_V \epsilon_L^{(KQ)} + B_{W2} L_\sigma B_{W1} \epsilon_L^{(V)}.\end{aligned} \tag{132}$$

We can bound $\left\|g_L^{(tf)}\left(X; W_{1:L}\right) - g_L^{(tf)}\left(X; \widehat{W}_{1:L}\right)\right\|_{2,\infty}$ till $\left\|g_1^{(tf)}\left(X; W_{1:1}\right) - g_1^{(tf)}\left(X; \widehat{W}_{1:1}\right)\right\|_{2,\infty}$ iteratively. Let

$$\alpha_l = \prod_{j=l+1}^{L} B_{W2} L_\sigma B_{W1} B_V \left(1 + 4B_{KQ}\right), \tag{133}$$

then

$$
\begin{aligned}
&\max_{t\in T}\left|g^{(tfw)}\big(X_t;W_{1:L+1},w\big)-g^{(tfw)}\big(X_t;\widehat{W}_{1:L+1},\widehat{w}\big)\right| \\
&\le \epsilon^{(w)}+B_w\sum_{l=2}^{L}\alpha_l\Big(\epsilon_l^{(W2)}+\epsilon_l^{(b2)}+B_{W2}L_\sigma\epsilon_l^{(W1)}+B_{W2}L_\sigma\epsilon_l^{(b1)} \\
&\qquad +2B_{W2}L_\sigma B_{W1}B_V\epsilon_l^{(KQ)}+B_{W2}L_\sigma B_{W1}\epsilon_l^{(V)}\Big) \\
&\quad +B_w\alpha_1\Big(\epsilon_1^{(W2)}+\epsilon_1^{(b2)}+B_{W2}L_\sigma\epsilon_1^{(W1)}+B_{W2}L_\sigma\epsilon_1^{(b1)} \\
&\qquad +2B_{W2}L_\sigma B_{W1}B_V\epsilon_1^{(KQ)}+B_{W2}L_\sigma B_{W1}\epsilon_1^{(V)}\Big).
\end{aligned}
\tag{134}
$$

Recall that $\epsilon_1^{(KQ)}$ has a different input bound than the rest. Define $\beta_l^{(2)}:=B_w\alpha_l$, $\beta_l^{(1)}:=B_w\alpha_l\cdot B_{W2}L_\sigma$, $\beta_l^{(KQ)}:=B_w\alpha_l\cdot 2B_{W2}L_\sigma B_{W1}B_V$, and $\beta_l^{(V)}:=B_w\alpha_l\cdot B_{W2}L_\sigma B_{W1}$. To get the covering number, we solve

$$
\begin{aligned}
\min_{\epsilon_1,\ldots,\epsilon_L}\ \epsilon^{(w)}+\sum_{l=2}^{L}&\left(\beta_l^{(2)}\left(\epsilon_l^{(W2)}+\epsilon_l^{(b2)}\right)+\beta_l^{(1)}\left(\epsilon_l^{(W1)}+\epsilon_l^{(b1)}\right)+\beta_l^{(KQ)}\epsilon_l^{(KQ)}+\beta_l^{(V)}\epsilon_l^{(V)}\right) \\
&+\left(\beta_1^{(2)}\left(\epsilon_1^{(W2)}+\epsilon_1^{(b2)}\right)+\beta_1^{(1)}\left(\epsilon_1^{(W1)}+\epsilon_1^{(b1)}\right)+\beta_1^{(KQ)}\epsilon_1^{(KQ)}+\beta_1^{(V)}\epsilon_1^{(V)}\right).
\end{aligned}
\tag{135}
$$

Let $i\in\mathcal{I}:=\{W,W1,b1,b2,W2,KQ,V\}$. Recall that we have log covering numbers in the form of C_1/ϵ^2 and C_{B_x}/ϵ^2 for the function class $\{x\mapsto Wx\mid x\in\mathcal{X},W\in\mathcal{W}\}$. For $\epsilon_l^{(b1)}$ and $\epsilon_l^{(b2)}$, we apply Maurey's specification (See Theorem 3 in Zhang (2002)). Therefore, summing across all parameter blocks, the total log covering number can be bounded as

$$
\log N_\infty\left(\mathcal{F},\epsilon,\{x_i\}_{i=1}^n,\|\cdot\|_2\right)\le\sum_l\sum_i\frac{C_l^{(i)}}{\left(\epsilon_l^{(i)}\right)^2},
\tag{136}
$$

To find the tightest possible bound, we optimally distribute the total error budget ϵ among the individual perturbations $\epsilon_l^{(i)}$. This leads to the optimization problem

$$
\min_{\{\epsilon_l^{(i)}>0\}}\sum_l\sum_i\frac{C_l^{(i)}}{\left(\epsilon_l^{(i)}\right)^2}\quad\text{subject to}\quad\sum_l\sum_i\beta_l^{(i)}\epsilon_l^{(i)}=\epsilon.
\tag{137}
$$

Applying Lemma G.1, we have

$$
\begin{aligned}
\eta&=C_1^{\frac{1}{3}}\sum_{l=2}^{L}\Big(2\beta_l^{(2)^{\frac{2}{3}}}+2\beta_l^{(1)^{\frac{2}{3}}}+\beta_l^{(KQ)^{\frac{2}{3}}}+\beta_l^{(V)^{\frac{2}{3}}}\Big), \\
\widetilde{\eta}&=C_{B_x}^{\frac{1}{3}}\left(\beta_1^{(KQ)}\right)+C_1^{\frac{1}{3}}\left(2\beta_1^{(2)^{\frac{2}{3}}}+2\beta_1^{(1)^{\frac{2}{3}}}+\beta_1^{(V)^{\frac{2}{3}}}\right)+C_1^{\frac{1}{3}}.
\end{aligned}
\tag{138}
$$

Then the covering number is bounded by

$$\frac{(\widetilde{\eta}+\eta)^3}{\epsilon^2}. \tag{139}$$

□

Lemma G.6. *Under Lemma G.5, and suppose we have the norm bounds* $\mathcal{W} = \{W \in \mathbb{R}^{k\times d} : \|W\|_{2,1} \le B_W\}$ *with* $B_W > 0$ *for each* $W_l^{(1)}$, $W_l^{(2)}$, $W_l^{(KQ)}$, $W_l^{(V)}$, w, *and* $\mathcal{B} = \{b \in \mathbb{R}^d : \|b\|_1 \le B_b\}$ *with* $B_b > 0$ *for* $b_l^{(1)}$ *and* $b_l^{(2)}$. *Let the maximum of all norm bounds be* B. *Let* $x_1, \dots, x_n \in \mathbb{R}^d$ *satisfy* $\|x_i\|_2 \le B_x$, *we have the log covering number of* $g_{L+1}^{(tfw)}(X, W_{1:L+1}, w)$ *as*

$$\frac{(\widetilde{\eta}+\eta)^3}{\epsilon^2}. \tag{140}$$

where

$$\begin{aligned}
\eta &= \sum_{l=2}^{L}\Big(B^2\log(dn)\Big)^{\frac{1}{3}}\Big(\beta_l^{(2)^{\frac{2}{3}}} + \beta_l^{(1)^{\frac{2}{3}}} + \beta_l^{(KQ)^{\frac{2}{3}}} + \beta_l^{(V)^{\frac{2}{3}}}\Big) + \big(B^2\log(2d+1)\big)^{\frac{1}{3}}\Big(\beta_l^{(2)^{\frac{2}{3}}} + \beta_l^{(1)^{\frac{2}{3}}}\Big) \\
\widetilde{\eta} &= \Big(B^2B_x^2\log(dn)\Big)^{\frac{1}{3}}\Big(\beta_1^{(KQ)}\Big) + \Big(B^2\log(dn)\Big)^{\frac{1}{3}}\Big(1 + \beta_1^{(2)^{\frac{2}{3}}} + \beta_1^{(1)^{\frac{2}{3}}} + \beta_1^{(V)^{\frac{2}{3}}}\Big) \\
&\quad + \big(B^2\log(2d+1)\big)^{\frac{1}{3}}\Big(\beta_1^{(2)^{\frac{2}{3}}} + \beta_1^{(1)^{\frac{2}{3}}}\Big),
\end{aligned} \tag{141}$$

with $\beta_l^{(2)} := B_w\alpha_l$, $\beta_l^{(1)} := B_w\alpha_l \cdot B_{W2}L_\sigma$, $\beta_l^{(KQ)} := B_w\alpha_l \cdot 2B_{W2}L_\sigma B_{W1}B_V$, *and* $\beta_l^{(V)} := B_w\alpha_l \cdot B_{W2}L_\sigma B_{W1}$ *and* $\alpha_l = \prod_{j=l+1}^{L} B_{W2}L_\sigma B_{W1}B_V\,(1+4B_{KQ})$.

Proof. By applying Lemma 4.6 from Edelman et al. (2022) and Maurey's specification in Theorem 3 of Zhang (2002) to Lemma G.5, we complete the proof. □

Lemma G.7 (Covering number of Transformer–MLP composition)**.** *Under the assumptions of Lemma G.5, let* $\mathcal{G}^{(TF)} = \big\{g_{L+1}^{(tfw)}(\cdot; W_{1:L+1}, w)\big\}$ *be the class of Transformer functions with parameters satisfying the stated norm bounds, and let* $\mathcal{F}^{(MLP)}$ *be the MLP class defined in Section 3.2 satisfying* $\max_i \|W_i\|_\infty \le B$ *and* $\max_i |b_i|_\infty \le B$ *for* $i = 0, \dots, D$ *and* $B > 0$. *Assume that* $\mathcal{F}^{(MLP)}$ *is uniformly Lipschitz with* $\sup_{f\in\mathcal{F}^{(MLP)}} \mathrm{Lip}(f) \le C$. *Define the composed class*

$$\mathcal{F} := \mathcal{F}^{(MLP)} \circ \mathcal{G}^{(TF)} = \{f \circ g : f \in \mathcal{F}^{(MLP)},\ g \in \mathcal{G}^{(TF)}\}. \tag{142}$$

Then, for any $\varepsilon > 0$, *the* $\|\cdot\|_\infty$*-covering number of* $\mathcal{F}$ *satisfies*

$$\log N_\infty(\varepsilon,\ \mathcal{F}) \le O(d_m \cdot \log\Big(16\,\varepsilon^{-1}D\,d^2\,{d_m}^{2D}\Big) + \frac{4C^2(\widetilde{\eta}+\eta)^3}{\varepsilon^2}. \tag{143}$$

where $\widetilde{\eta}$ and η are defined as in Lemma G.5.

Proof. For notational simplicity, denote the Transformer block by $g := g_{L+1}^{(tfw)}(X; W_{1:L+1}, w)$, and the MLP network by $f := f^{(mlp)}$. We consider the composed function class

$$\mathcal{F} := \{f \circ g : f \in \mathcal{F}^{(MLP)},\ g \in \mathcal{G}^{(TF)}\}. \tag{144}$$

Since $f \in \mathcal{F}^{(MLP)}$ is Lipschitz continuous, for any $g, g' \in \mathcal{G}^{(TF)}$ we have

$$\|f \circ g - f \circ g'\|_\infty \leq \mathrm{Lip}(f)\, \|g - g'\|_\infty. \tag{145}$$

Under the parameter constraints of Schmidt-Hieber (2020), the MLP class is uniformly Lipschitz; that is,

$$\sup_{f \in \mathcal{F}^{(MLP)}} \mathrm{Lip}(f) \leq C < \infty. \tag{146}$$

Hence, the map $g \mapsto f \circ g$ is C-Lipschitz uniformly over $f \in \mathcal{F}^{(MLP)}$.

By Lemma 3.1 in Petrova and Wojtaszczyk (2023), which controls covering numbers under Lipschitz mappings, it follows that

$$N_\infty(\varepsilon,\ f \circ \mathcal{G}^{(TF)}) \ \leq\ N_\infty(\varepsilon/C,\ \mathcal{G}^{(TF)}). \tag{147}$$

To cover the entire composed class $\mathcal{F}$, we discretize both the MLP and Transformer classes. Therefore

$$N_\infty(\varepsilon,\ \mathcal{F}) \ \leq\ N_\infty(\varepsilon/2,\ \mathcal{F}^{(MLP)}) \cdot N_\infty(\varepsilon/(2C),\ \mathcal{G}^{(TF)}), \tag{148}$$

and further,

$$\log N_\infty(\varepsilon,\ \mathcal{F}) \ \leq\ \log N_\infty(\varepsilon/2,\ \mathcal{F}^{(MLP)}) + \log N_\infty(\varepsilon/(2C),\ \mathcal{G}^{(TF)}). \tag{149}$$

MLP covering number. Define that $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ for two sequences. Let $\lceil \cdot \rceil$ denotes the ceiling function and $\vee$ denotes the maximum operator. Recalling the output MLP neural network in (12). By Lemma 5 of Schmidt-Hieber (2020), for any integers $m \geq 1$, $N \geq (\beta+1)^s \vee (L_{\mathcal{I}}+1)\exp(s)$ and any $\delta > 0$, we have $\mathcal{F}^{(MLP)}(D, \widetilde{d}_m)$ with number of parameters $S^{(net)}$ such that

$$\log N_\infty(\delta,\ \mathcal{F}^{(MLP)}) \leq (S^{(net)}+1) \log\left(2\,\delta^{-1}\,(D+1)V^2\right) \tag{150}$$

where $V := 2(d+1) \cdot \prod_{i=1}^{D-1} \left(d_m^{(i)} + 1\right)$, D is the depth, $d_m^{(i)}$ the dimension of each hidden layer. Applying (150) with $\delta = \varepsilon/2$ and consider the defined MLP in section 3.2 gives

$$\log N_\infty(\varepsilon/2,\ \mathcal{F}^{(MLP)}) \leq (S^{(net)} + 1) \log\left(4\,\varepsilon^{-1}\,(D+1)V^2\right). \tag{151}$$

Since $D = 8 + (m+5)\,(1 + \lceil \log_2(s \vee \beta) \rceil)$, $d_m := d_m^{(i)} = 6(s + \lceil \beta \rceil)N$ and $S^{(net)} \leq 141(s+\beta+1)^{3+s}N(m+6)$ by Theorem 5 in Schmidt-Hieber (2020), combining the above result for MLP and choosing $m \asymp \log d_m$, we have

$$\log N_\infty(\varepsilon/2,\ \mathcal{F}^{(MLP)}) = O\left(\left(d_m \log(d_m) + 1\right) \cdot \log\left(16\,\varepsilon^{-1}\,(D+1)(d+1)^2(d_m+1)^{2D-2}\right)\right). \tag{152}$$

Transformer covering number. By Lemma G.5 (stated above), for any $\delta > 0$,

$$\log N_\infty(\delta,\ \mathcal{G}^{(TF)}) \leq \frac{(\widetilde{\eta} + \eta)^3}{\delta^2}. \tag{153}$$

Applying (153) with $\delta = \varepsilon/(2C)$ yields

$$\log N_\infty(\varepsilon/(2C),\ \mathcal{G}^{(TF)}) \leq \frac{(\widetilde{\eta} + \eta)^3}{(\varepsilon/(2C))^2} = \frac{4C^2(\widetilde{\eta} + \eta)^3}{\varepsilon^2}. \tag{154}$$

Combining both covering numbers by substituting (151) and (154) into (149) gives

$$\log N_\infty(\varepsilon,\ \mathcal{F}) \leq O\left(d_m \log(d_m) \cdot \log\left(\varepsilon^{-1}\,D\,d\,d_m^D\right)\right) + \frac{4C^2(\widetilde{\eta} + \eta)^3}{\varepsilon^2}. \tag{155}$$

□

H Bias Rate

We follow the setting of Edelman et al. (2022) and slightly adapt their definitions to align with our model architecture. We begin with the following definitions.

1. ($\mathcal{I}$**-sparsity**) A function $f : [0,1]^n \mapsto \mathcal{Y}$ is $\mathcal{I}$-sparse if it only depends on a fixed subset $\mathcal{I} \subseteq \{1, \cdots, n\}$ of its inputs:

$$x_i = x_i' \quad \forall i \in \mathcal{I} \quad \Longrightarrow \quad f(x) = f(x'). \tag{156}$$

2. (γ**-injective**) An $\mathcal{I}$-sparse function f is γ-injective if there exists $\gamma > 0$ such that

$$\|f(x) - f(x')\|_\infty \;\geq\; \gamma \|\Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x'\|_\infty, \qquad \forall x, x' \in [0,1]^n, \tag{157}$$

   where $\Pi_{\mathcal{I}}$ is a projection onto coordinates indexed by $\mathcal{I}$.

We then make the following assumptions.

Assumption H.1 (Target Function)**.** *Define $x \in [0,1]^n$ and let $\mathbb{Z} : [0,1]^n \to [0,1]^{d \times n}$ be a fixed embedding with $Z = \mathbb{Z}(x)$. Assume there exists $\mathcal{I} \subseteq \{1, \ldots, n\}$ with $|\mathcal{I}| = s \leq \min\{n, d\}$, and a β-Hölder function $\phi : [0,1]^s \to \mathbb{R}$ with constant $L_{\mathcal{I}}$, such that the target function $g^* : [0,1]^{d \times n} \to \mathbb{R}$ satisfies, for all $x \in [0,1]^n$,*

$$g^* \circ \mathbb{Z}(x) = \widetilde{\phi} \circ \widetilde{g}_{\mathcal{I}}(x), \tag{158}$$

where $\widetilde{g}_{\mathcal{I}}(x)$ is $\mathcal{I}$-sparse, 1-bounded and γ-injective, and $\widetilde{\phi} : [0,1]^d \to \mathbb{R}$ is defined by

$$\widetilde{\phi}(u) := \phi\left((u_1, \ldots, u_s)\right). \tag{159}$$

Assumption H.2 (Recoverability by Self-Attention)**.** *Assume $\mathbb{Z}$ satisfies structural conditions similar to those in Appendix B of Edelman et al. (2022) (e.g., approximate orthogonality of positional encodings). Specifically, the embedding $Z = \mathbb{Z}(x)$ satisfies the column-wise decomposition*

$$Z_{:,j} = p_j + x_j u, \qquad j \in [n], \tag{160}$$

where $\|p_j\|_2 = \|u\|_2 = 1$, $\langle p_j, u \rangle = 0$, and $|\langle p_i, p_j \rangle| \leq \Delta$ for $i \neq j$.

Assumption H.2 essentially requires that the key latent information be appropriately embedded in the input

matrix so that attention layers can extract it effectively, as shown later. Following Edelman et al. (2022), we decompose the input columns into two components, and for illustrative purposes, we adopt a specific form, while acknowledging that other forms may also be viable[15].

Lemma H.1. *Let $x \in [0,1]^n$ and $\mathbb{Z} : [0,1]^n \to \mathbb{R}^{d\times n}$ with $Z = \mathbb{Z}(x)$ be a fixed embedding satisfying the structural conditions in Assumption H.2, with approximate orthogonality parameter Δ and $\Delta < 1/s$. Under Assumption H.1, there exists $g : \mathbb{R}^{d\times n} \mapsto \mathbb{R}$ in the Transformer function class $\mathcal{G}$ defined in* (13) *such that g approximates g^* under $\mathbb{Z}$, i.e.,*

$$\sup_{Z\in\mathbb{Z}([0,1]^n)} \left|g(Z) - g^*(Z)\right| \;\le\; \varepsilon. \tag{161}$$

And the weights satisfy

$$\|W_h^{(Q)}\|_{2,1} \le \frac{\log\left(\frac{4n}{\gamma}\right)}{1-2\Delta}, \qquad \|W_h^{(K)}\|_{2,1} \le 1, \qquad \|W_h^{(V)}\|_{2,1} \le 1, \qquad \|W_h^{(O)}\|_{2,1} \le 1 \tag{162}$$

for $h = 1,\dots,s$. And the weights in MLP satisfy $\max_i \|W_i\|_\infty \le B$ and $\max_i |b_i|_\infty \le B$ for $i = 0,\dots,D$ and $B > 0$.

Proof. Fix $\mathcal{I} = \{i_1,\dots,i_s\} \subset [n]$ and let $Z = \mathbb{Z}(x)$.

Step 1: (Attention Approximation Error). Taking $x \in [0,1]^n$, we aim to approximate the 1-bounded γ-injective (e.g., with $\gamma = 1$) $\mathcal{I}$-sparse function

$$g_{\mathcal{I}}(x) := \sum_{i\in\mathcal{I}} x_i\, e^{(d)}_{\pi(i)} \qquad \in \mathbb{R}^d, \tag{163}$$

where $\pi : \mathcal{I} \to [s]$ is a bijection, and we write $i_h := \pi^{-1}(h)$.

Set r the index of the readout column. Under Assumption H.2, we define

$$W_h^{(Q)} := R\, e_1^{(k)} p_r^\top, \quad W_h^{(K)} := e_1^{(k)} p_{i_h}^\top, \quad W_h^{(V)} := e_1^{(k)} u^\top, \quad W_h^{(O)} := e_h^{(d)} {e_1^{(k)}}^\top, \tag{164}$$

with $R > 0$ to be specified. For head h,

$$\big(\alpha_h(Z)\big)_{:,r} := \sigma_S\left(Z^\top {W_h^{(K)}}^\top W_h^{(Q)} Z_{:,r}\right) \qquad \in [0,1]^n \tag{165}$$

[15]One form considered Edelman et al. (2022) is that $Z_{:,j} = p_j + x_j u_1 + (1-x_j)u_0$ with $j \in [n]$, where x_j is binary and u_1, u_0 are orthogonal to each other. Note that our proof also extends straightforwardly to such a case.

By Lemma B.2 and B.7 in Edelman et al. (2022), we have

$$\left\|\left(\alpha_h(Z)\right)_{:,r} - e_{i_h}^{(n)}\right\|_1 \leq \frac{2n}{\exp(R(1-2\Delta))}. \tag{166}$$

Since $W_h^{(V)} Z = e_1^{(k)}(u^\top Z) = e_1^{(k)}(x_1, \ldots, x_n) = e_1^{(k)} \sum_{j=1}^n x_j \left(e_j^{(n)}\right)^\top$, then

$$W_h^{(O)} W_h^{(V)} Z = e_h^{(d)} {e_1^{(k)}}^\top e_1^{(k)} u^\top Z = e_h^{(d)} \sum_{j=1}^n x_j \, (e_j^{(n)})^\top. \tag{167}$$

The attention output for head h is

$$g_h^{(msa)}(\mathbb{Z}(x)) = e_h^{(d)} \sum_{j=1}^n x_j {e_j^{(n)}}^\top \left(\alpha_h(\mathbb{Z}(x))\right)_{:,r} = e_h^{(d)} \sum_{j=1}^n x_j \left(\alpha_h(\mathbb{Z}(x))\right)_{j,r} \qquad \in [0,1]^d. \tag{168}$$

Summing over $h = 1, \ldots, s$ gives the MSA output

$$g^{(msa)}(\mathbb{Z}(x)) := \sum_{h=1}^s f_h^{(msa)}(\mathbb{Z}(x)) = \sum_{h=1}^s e_h^{(d)} \sum_{j=1}^n x_j \left(\alpha_h(\mathbb{Z}(x))\right)_{j,r} \qquad \in [0,1]^d. \tag{169}$$

After the MSA sublayer, the FFN sublayer $\mathcal{G}^{(FF)}$ acts column-wise. We may choose the FFN parameters such that the overall FFN mapping acts as the identity on the readout. As a remark, the FFN can also be used to eliminate the residual connection (see Lemma 5 in Jiao et al. (2025)). In what follows, we take $g^{(tf)}(\mathbb{Z}(x)) := g^{(msa)}(\mathbb{Z}(x))$. Thus

$$\begin{aligned}
\left\|g^{(tf)}(\mathbb{Z}(x)) - g_{\mathcal{I}}(x)\right\|_\infty &= \max_{h\in[s]} \left| \sum_{j=1}^n x_j \left(\alpha_h(Z)\right)_{j,r} - x_{i_h} \right| \\
&= \max_{h\in[s]} \left| x^\top \left(\left(\alpha_h(Z)\right)_{:,r} - e_{i_h}^{(n)} \right) \right| \\
&\leq \max_{h\in[s]} \|x\|_\infty \left\| \left(\alpha_h(Z)\right)_{:,r} - e_{i_h}^{(n)} \right\|_1 \\
&\leq \max_{h\in[s]} \left\| \left(\alpha_h(Z)\right)_{:,r} - e_{i_h}^{(n)} \right\|_1 \\
&\leq \frac{2n}{\exp\big(R(1-2\Delta)\big)}.
\end{aligned} \tag{170}$$

Let $x, x' \in [0,1]^n$ satisfy $\Pi_{\mathcal{I}} x \neq \Pi_{\mathcal{I}} x'$. Set $Z := \mathbb{Z}(x)$ and $Z' := \mathbb{Z}(x')$. For each $h \in [s]$, note that the attention weights $\left(\alpha_h(Z)\right)_{:,r}$ depend only on the positional encodings $\{p_j\}_{j=1}^n$ (and the readout index r),

hence they are independent of x. Moreover, by the explicit form of the construction,

$$e_h^{(d)} \cdot g^{(tf)}(Z) = \sum_{j=1}^{n} x_j \left(\alpha_h(Z)\right)_{j,r}, \qquad e_h^{(d)} \cdot g_{\mathcal{I}}(x) = x_{i_h}. \tag{171}$$

we have for each $h \in [s]$,

$$\begin{aligned}
&\left| e_h^{(d)} \cdot \left(g^{(tf)}(Z) - g^{(tf)}(Z')\right) - e_h^{(d)} \cdot \left(g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x')\right) \right| \\
&= \left| \sum_{j=1}^{n} \left(\alpha_h(Z)\right)_{j,r} (x_j - x'_j) - (x_{i_h} - x'_{i_h}) \right| \\
&= \left| \sum_{j=1}^{n} \left(\left(\alpha_h(Z)\right)_{j,r} - (e_{i_h}^{(n)})_j \right) (x_j - x'_j) \right| \\
&= \left| \sum_{j \in \mathcal{I}} \left(\left(\alpha_h(Z)\right)_{j,r} - (e_{i_h}^{(n)})_j \right) (x_j - x'_j) \right| \\
&\leq \left\| \left(\alpha_h(Z)\right)_{:,r} - e_{i_h}^{(n)} \right\|_1 \cdot \left\| \Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x' \right\|_\infty \\
&\leq \delta \left\| \Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x' \right\|_\infty .
\end{aligned} \tag{172}$$

Taking the maximum over $h \in [s]$ yields

$$\sup_{\substack{x,x' \\ \Pi_{\mathcal{I}} x \neq \Pi_{\mathcal{I}} x'}} \frac{\left\| g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}(\mathbb{Z}(x')) - \left(g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x')\right) \right\|_\infty}{\left\| \Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x' \right\|_\infty} \leq \delta. \tag{173}$$

Now use the increment control to get injectivity of the Transformer representation. For any x, x' with $\Pi_{\mathcal{I}} x \neq \Pi_{\mathcal{I}} x'$,

$$\begin{aligned}
& \left\| g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}\left(\mathbb{Z}\left(x'\right)\right) \right\|_\infty \\
= \;& \left\| \left(g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x')\right) + \left(g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}(\mathbb{Z}(x')) - \left(g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x')\right)\right) \right\|_\infty \\
\geq \;& \left\| g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x') \right\|_\infty - \left\| g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}(\mathbb{Z}(x')) - \left(g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x')\right) \right\|_\infty \\
\geq \;& \left\| g_{\mathcal{I}}(x) - g_{\mathcal{I}}(x') \right\|_\infty - \delta \left\| \Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x' \right\|_\infty .
\end{aligned} \tag{174}$$

Since $g_{\mathcal{I}}$ is γ-injective under Assumption H.1,

$$\left\| g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}\left(\mathbb{Z}\left(x'\right)\right) \right\|_\infty \geq (\gamma - \delta) \left\| \Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x' \right\|_\infty . \tag{175}$$

Choose R such that $\delta \le \gamma/2$, i.e.

$$R \ge \frac{\log(4n/\gamma)}{1-2\Delta} \tag{176}$$

Then $g^{(tf)} \circ \mathbb{Z}$ is $\gamma/2$-injective w.r.t. $\Pi_{\mathcal{I}} x$.

Define the set

$$\mathcal{S} := \left\{ g^{(tf)}(\mathbb{Z}(x)) : x \in [0,1]^n \right\} \subset \mathbb{R}^d. \tag{177}$$

and define a recovery map

$$\mathrm{Rec} : \mathcal{S} \to [0,1]^s, \qquad \mathrm{Rec}\left(g^{(tf)}(\mathbb{Z}(x)) \right) := \Pi_{\mathcal{I}} x. \tag{178}$$

Since (175) implies that Rec is Lipschitz on $\mathcal{S}$: for any x, x',

$$\begin{aligned} \left\| \mathrm{Rec}\left(g^{(tf)}(\mathbb{Z}(x)) \right) - \mathrm{Rec}\left(g^{(tf)}(\mathbb{Z}(x')) \right) \right\|_\infty &= \|\Pi_{\mathcal{I}} x - \Pi_{\mathcal{I}} x'\|_\infty \\ &\le \frac{1}{\gamma - \delta} \|g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}(\mathbb{Z}(x'))\|_\infty \\ &\le \frac{2}{\gamma} \|g^{(tf)}(\mathbb{Z}(x)) - g^{(tf)}(\mathbb{Z}(x'))\|_\infty. \end{aligned} \tag{179}$$

Step 2: (MLP Approximation Error). For any $u, v \in \mathcal{S}$, using the β-Hölder continuity of ϕ and (179),

$$\left| \phi(\mathrm{Rec}(u)) - \phi(\mathrm{Rec}(v)) \right| \le L_{\mathcal{I}} \left(\frac{2}{\gamma} \right)^\beta \|u - v\|_\infty^\beta. \tag{180}$$

Thus the function

$$u \;\mapsto\; \phi(\mathrm{Rec}(u)), \qquad u \in \mathcal{S}, \tag{181}$$

is β-Hölder on $\mathcal{S}$ with constant $L_{\mathcal{I}}(2/\gamma)^\beta$ and intrinsic dimension s.

Recalling the output MLP neural network in (12). Define that $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ for two sequences. Let $\lceil \cdot \rceil$ denotes the ceiling function and $\vee$ denotes the maximum operator. By Schmidt-Hieber's Theorem 5, there exists an MLP $f^{(mlp)}$ with depth $D = 8 + (m+5)\left(1 + \lceil \log_2(s \vee \beta) \rceil\right)$, hidden layer width $d_m := d_m^{(i)} = 6(s + \lceil \beta \rceil)N$ for $i = 1, \ldots, D-1$, and number of nonzero parameters $S^{(net)} \le C_0 \, (s+\beta+1)^{3+s} \, N(m+6)$, where m and N are any integers satisfying $m \ge 1$ and $N \ge (\beta+1)^s \vee (L_{\mathcal{I}}+1)\exp(s)$ such that

$$\sup_{u \in \mathcal{S}} \left| g^{(mlp)}(u) - \phi(\mathrm{Rec}(u)) \right| = O\left(L_{\mathcal{I}} \left(\frac{2}{\gamma} \right)^\beta \left(N 2^{-m} + N^{-\beta/s} \right) \right). \tag{182}$$

Since $\mathrm{Rec}(f^{(tf)}(\mathbb{Z}(x))) = \Pi_{\mathcal{I}} x$, for all $x \in [0,1]^n$,

$$g^{(mlp)}\left(g^{(tf)}(\mathbb{Z}(x))\right) = \phi(\Pi_{\mathcal{I}} x) = f^*_{\mathcal{I}}(x). \tag{183}$$

Noting that $d_m^{(i)} \asymp N$, we may equivalently write the bias term as

$$\sup_{x\in[0,1]^n} \left| g^{(mlp)}\left(g^{(tf)}(\mathbb{Z}(x))\right) - g^*(x) \right| \lesssim \left(\frac{2}{\gamma}\right)^{\beta} \left(d_m\, 2^{-m} + d_m^{-\beta/s} \right). \tag{184}$$

We choose $m \asymp \log N \asymp \log d_m$, so that $d_m 2^{-m} \lesssim d_m^{-\beta/s}$ and hence

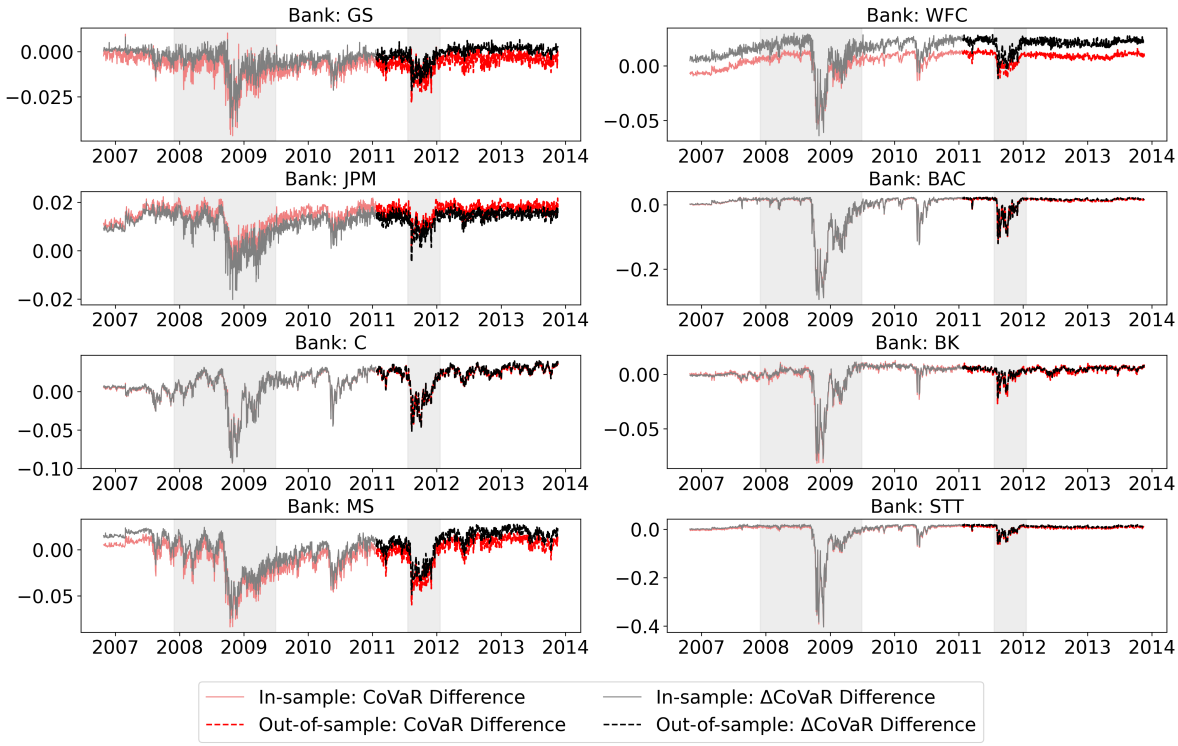
$$\sup_{x\in[0,1]^n} \left| g^{(mlp)}\left(g^{(tf)}(\mathbb{Z}(x))\right) - g^*(x) \right| = O\left(\left(\frac{2}{\gamma}\right)^{\beta} \left(d_m\right)^{-\beta/s} \right), \tag{185}$$

under the choice of R in (176).

□

I Additional Figure(s)

Figure 15: CoVaR and ΔCoVaR Differences: Text vs. No-Text (Eight G-SIBs)



Notes: Differences in CoVaR (red) and ΔCoVaR (black) between models with and without text for the eight Global Systemically Important Banks.