Forecasting Credit Ratings: A Case Study where Traditional Methods Outperform Generative LLMs

Felix Drinkall*, Janet B. Pierrehumbert*‡, Stefan Zohren*†
*Department of Engineering Science, University of Oxford
†The Alan Turing Institute
‡Faculty of Linguistics, University of Oxford
felix.drinkall@eng.ox.ac.uk

Abstract

Large Language Models (LLMs) have been shown to perform well for many downstream tasks. Transfer learning can enable LLMs to acquire skills that were not targeted during pre-training. In financial contexts, LLMs can sometimes beat well-established benchmarks. This paper investigates how well LLMs perform at forecasting corporate credit ratings. We show that while LLMs are very good at encoding textual information, traditional methods are still very competitive when it comes to encoding numeric and multimodal data. For our task, current LLMs perform worse than a more traditional XGBoost architecture that combines fundamental and macroeconomic data with high-density text-based embedding features. We investigate the degree to which the text encoding methodology affects performance and interpretability. The dataset reconstruction and model code from this paper is provided[1].

1 Introduction

Corporate credit ratings indicate a borrower's ability to service its debt obligations and are a forward-looking measure of a company's health (Baresa et al., 2012). A company's credit rating is significant since it affects the cost of raising capital, which in turn could finance future infrastructure to increase revenue or profitability. An optimistic rating can result in a virtuous cycle whereby it is easier to raise money and grow the business (Cho et al., 2020), and a pessimistic rating can result in a vicious cycle in which competition can grow faster due to cheaper debt obligations. Knowing which cycle a company may enter can be advantageous to investors. Many major funds are also not allowed to own sub-prime assets, which makes forecasting a drop in credit rating very important so that the fund has more time to divest from the asset, which could result in a higher close price.
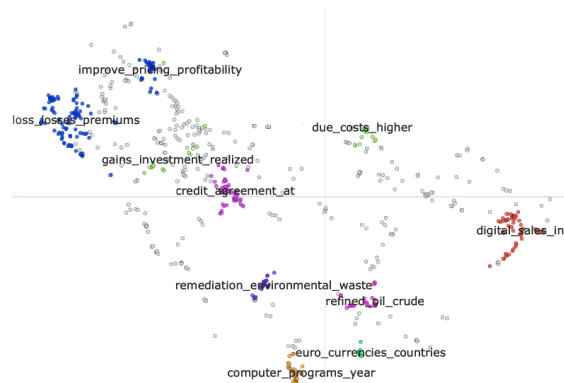

Figure 1: Example of the best-performing feature - high-density clustering (Drinkall et al., 2022). Each dot represents a sentence, and the colored areas representing high-density regions of the embedding space.

Recently, there has been a surge of interest in text-based forecasting (Xu and Cohen, 2018; Yang et al., 2020; Nie et al., 2024). One reason for this trend is the progress that has been made in text modelling in general (Zoph et al., 2022; Touvron et al., 2023). Given that financial news is often first disseminated through written or spoken communications (Boulland et al., 2016), rather than in numeric or tabular formats, there has been a hope that important information can be included in models sooner than was possible without using linguistic information. Another reason is that language can provide relevant context and forward-looking information, whereas financial numeric reporting alone is inherently retrospective. Contained within a company's filings, text can provide insights about the future strategic direction of the company as well as historical information.

The majority of text-based forecasting research has been focused on short text sequences, drawing primarily from sources such as social media (Xu and Cohen, 2018), news articles (Zhang et al., 2018), and analyst recommendations (Rekabsaz et al., 2017). In contrast, many fundamental fi-

---

[1] https://github.com/FelixDrinkall/credit-ratings-project

nancial documents, such as company filings, earnings call transcripts, and patents, are very long. Considering that the shorter texts often serve as summaries, reflections or commentaries on the detailed primary sources and that speed of information acquisition is essential in finance (Rzayev et al., 2023), there must be more focus on text-based forecasting for longer text sequences. This paper evaluates the most effective ways to model longer text sequences within a text-based forecasting task.

Linked to the recent progress of LLMs, there has been growing interest in applying these models in a variety of downstream applications (Kaddour et al., 2023). It has been shown that as generative LLMs scale, they acquire abilities that were not present in smaller LLM variants (Zoph et al., 2022), such as modular arithmetic (Srivastava et al., 2023), NLU (Hendrycks et al., 2021), commonsense reasoning (Lin et al., 2022), fact-checking (Rae et al., 2022) and so on. These abilities have been impressive, but it is best not to be over-optimistic. The lack of training data transparency associated with some of the best-performing LLMs means that we cannot be certain whether some of the performance gains are due to the memorisation of benchmarks being in the training datasets (Bender et al., 2021; Sainz et al., 2023; Xu et al., 2024; Balloccu et al., 2024). Generative LLMs also seem to have a mediocre understanding of concepts like negation and complex logical reasoning (Kassner and Schütze, 2020; Lorge and Pierrehumbert, 2023; Huang and Chang, 2023; Truong et al., 2023). These limitations in the capabilities of LLMs could prove to be very consequential in financial contexts. In this paper, we test generative LLMs on a complex linguistic task, which has never been fully solved by human experts: credit rating forecasting. We show that while LLMs encode text-based information very well, they are not good at incorporating numeric information, and underperform a boosting-tree baseline.

The contributions of this paper are as follows:

- We show that generative LLMs are poor at encoding numerical information, and underperform traditional methods.

- To our knowledge, this is the first use of modern language modelling techniques in a credit rating forecasting task.

- A financial dataset that can be reproduced with an academic WRDS licence.

- A benchmark of techniques for encoding long-sequence text in a forecasting task.

2 Related Work

2.1 Text-based forecasting

2.1.1 Encoding Text for Forecasting

The predominant approach in text-based forecasting has focused on the extraction of interpretable features like sentiment and uncertainty scores (Song and Shin, 2019; An et al., 2023). Rule-based sentiment using diverse lexicons has dominated the literature (Mohammad, 2020; Kalamara et al., 2022; Barbaglia et al., 2023). Lexicons tailored to specific domains generally surpass broader lexicons in predictive tasks (Loughran and McDonald, 2011; Li et al., 2014). Nevertheless, lexicons overlook contextual nuance and inadequately address common linguistic phenomena like negation. To mitigate these limitations, efforts have been made to integrate more sophisticated sentiment classifiers (An et al., 2023; Ayyappa et al., 2023). However, sentiment presupposes that important information can be encapsulated within a single dimension. To avoid an overly simple and prescriptive feature set, unsupervised methods have been used in feature exploration: TF-IDF (Jones, 1972), Latent Dirichlet Allocation (LDA) (Wang et al., 2017; Kanungsukkasem and Leelanupab, 2019). However, the arrival of contemporary topic models has gradually eclipsed LDA, fostering the adoption of transformer-derived topic models into forecasting tasks (Drinkall et al., 2022).

Recently, some studies have used the representations from encoder-based LLMs as features for text-based forecasting. LLMs exploit high-dimensional embeddings to capture the linguistic meaning of words (Devlin et al., 2019; Radford et al., 2019) and sentences (Reimers and Gurevych, 2019), with representation dimensionality ranging from 384 (Wang et al., 2020) to 5192 (Touvron et al., 2023). Such dimensionality poses challenges when used with smaller datasets. Nonetheless, there has been some success in incorporating these methods into text-based forecasting (Sawhney et al., 2020; Lee et al., 2023). However, the effectiveness of LLMs is often hampered by their limited context windows. Recent advancements have seen an increase in context window sizes, thanks in part to better GPU infrastructure making it computationally feasible, the implementation of attention sparsification techniques (Tay et al., 2022), and positional encoding

hacks (Chen et al., 2023). There are also methods that combine the use of transformer-based LLMs with feature-based methods, such as topic clusters (Grootendorst, 2022; Drinkall et al., 2022), or emotions (Liapis and Kotsiantis, 2023).

2.1.2 Generative Multimodal Forecasting

In addition to the adoption of encoder-based LLMs like BERT (Devlin et al., 2019), generative LLMs have been used in text-based forecasting tasks. Generative LLMs use masked-self attention to model text in an autoregressive manner. The GPT (Radford et al., 2019) and Llama (Touvron et al., 2023) model families are part of the generative model class. LLMs have been used as a backbone model for generative time-series forecasting models (Cao et al., 2023; Chang et al., 2023; Zhou et al., 2024; Liu et al., 2024), showing that an adapted generative language model can forecast weather, electricity and several other domains without relying on traditional text inputs. Liu et al. 2024 used eight text-based frames in order to create a general time-series modal that could be applied to several domains, and in so doing encoded both text and numerical information in a GPT-2-small model to generate the predicted time-series.

Beyond the use of text-based frames in generative forecasting tasks, GPT4MTS (Lee et al., 2023) encoded both news and time-series information before passing the concatenated input sequence through a pre-trained GPT-2 model. FinMA (Xie et al., 2024) and PromptCast (Xue and Salim, 2024) evaluated the performance of LLMs on stock movement prediction by converting the time-series information into natural language and prompting the language model for the predicted direction. Yu et al. 2023 takes this further by passing exclusively text information into the prompt for a financial forecasting task. There has been little comparison between these generative methods and the more traditional discriminative methods when applied to multimodal information.

2.2 Credit Rating Prediction

Research in **C**redit **R**ating **P**rediction (CRP) has tended to focus on predicting the absolute credit rating at time $t = 0$ given the feature set $F_{t=0}$ (Li et al., 2023; Galil et al., 2023; Tavakoli et al., 2023). This approach takes the perspective of the rating agencies and is useful for identifying anomalies where the existing credit rating classification appears to be implausible or inconsistent with cur-

rent financial indicators (Lokanan et al., 2019). However, predicting the absolute rating level is more simple and not as useful as predicting a future change. There is some limited research on **C**redit **R**ating **F**orecasting (CRF), where the target is the movement direction of the credit rating at time $t = 1$. This task takes the perspective of the investor seeking to predict whether an asset is likely to be classified as more or less risky in the next time period, and is the task outlined in this paper.

There have been some attempts to incorporate linguistic information into both corporate risk (Fei et al., 2015; Cao et al., 2024) and default prediction (Mai et al., 2019; Stevenson et al., 2021). Some papers have shown how text can help improve consumer credit lending (Hurley and Adebayo, 2016; Babaei and Giudici, 2024). There has also been some attempts to include textual data in CRP and CRF tasks (Chen and Chen, 2022; Muñoz-Izquierdo et al., 2022; Tavakoli et al., 2023). The majority of the existing literature uses lexicons, keywords or sentiment to encode the text (Kogan et al., 2009; Fei et al., 2015; Mai et al., 2019; Muñoz-Izquierdo et al., 2022; Chen and Chen, 2022). There have been some studies that have utilized encoder-based LLM representations (Stevenson et al., 2021; Tavakoli et al., 2023; Cao et al., 2024). There has been some work exploring how well generative models perform at assessing credit lending applications (Babaei and Giudici, 2024), and value at risk in general (Cao et al., 2024), but there has been no work benchmarking how well modern generative LLMs perform on a CRF task. Understanding generative LLMs' relative strengths relative to more traditional methods is an important contribution to the existing literature.

3 Dataset

In part due to the lack of large open-source or readily available datasets with temporal metadata, most of the financial text-based forecasting studies have either focused on expensive proprietary datasets, or datasets spanning 2-3 years (Xu and Cohen, 2018; Soun et al., 2022), making results hard to replicate and potentially biased to a specific time. While temporal bias in language-based tasks is hard to avoid due to limited historical data (Drinkall et al., 2024), we aim to reduce this by using a dataset spanning 23 years, increasing the models' exposure to different economic contexts. The cost and

lack of transparency of large datasets have hindered progress in the field and made it harder to build on promising work due to the difficulty of replicating results. As such, all data used in this paper is either open source or available with a WRDS subscription to enable effective dataset reconstruction. The data used in this paper is from US-based companies.

3.0.1 Credit ratings (C)

For the credit ratings, we used the Compustat Capital IQ dataset[2], using Standard & Poors' (S&P) ratings. These ratings cover the period from 1978 to 2017. S&P routinely assesses and assigns credit ratings to companies. Our paper predicts changes to the long-term credit ratings. Notably, we incorporate historical ratings from preceding quarters into our prediction models, acknowledging the distinct implications of a top-rated company (AAA) being downgraded compared to a lower-rated one (CC) experiencing a similar decline.

3.0.2 SEC filings

This paper uses 10-Q and 10-K filings available in the SEC's EDGAR database[3] to provide both textual context. They were chosen for their consistent structure which aids homogenous feature extraction. While most of the content in these filings is comprised of indexing, tables, and introductory text, we're interested in the parts that offer insights into a company's future financial health. As such, we've focused on the Management's Discussion and Analysis of Financial Condition and Results of Operations (MDA) section. We extract the MDA sections from all SEC filings - using the SEC-API[4] - for which we had credit rating data, spanning from Q1-1994 to Q2-2017. The API returns cleaned text, but we clean the text further by removing the remaining HTML, links and excessive spaces.

3.0.3 Fundamental data (F)

S&P emphasizes two components in their credit rating methodology: the financial and business risk profiles (Gillmor, 2015). While the text from the MDA section provides some insight into the qualitative business risk profile, numerical fundamental data is important to assess the financial health of a company. For the fundamental data, we use the Compustat Quarterly Fundamentals dataset[5]. The variables selected are outlined in Appendix

C. These variables were consistently reported for all the companies under consideration. Ideally, we would incorporate a broader range of fundamental variables, but expanding the variable set would result in fewer samples with complete data, thus limiting the scope of our analysis.

3.0.4 Macroeconomic data (M)

Adverse events in the world economy can also impact a company's ability to repay its debt. Many external forces can affect a company's future creditworthiness, however, we have identified three key areas from prior research in the area (Carling et al., 2007; Taylor et al., 2021): labour statistics, interest rates and foreign exchange data. For the labour statistics, we used the Bureau of Labour Statistics dataset[6]. For the interest rate and foreign exchange data, we used the Federal Reserve Bank Reports[7,8].

3.1 Dataset Construction

To maintain consistent periodicity in SEC filings, all data is aligned quarterly. The dataset spans from Q1 1994, when the SEC began electronic processing of filings, to Q2 2017, the last period with credit rating data from Compustat Capital IQ. Companies with incomplete records were excluded. As a result, when the number of lagged quarters used in the task is increased, the number of valid samples diminishes. This reduction is due to the lower probability of having complete data across many consecutive quarters, compared to when only the most recent quarter is considered.

Credit rating data is highly imbalanced, with 93.4% of companies maintaining the same score. While oversampling techniques like SMOTE (Chawla et al., 2002) are common for credit rating prediction (Pamuk and Schumann, 2023; Wang, 2022; Zhao et al., 2024), their application to text embeddings lacks consensus. To address this, we balanced the classes, reducing the dataset size. Training data spans Q1 1994 to Q4 2012, validation from Q1 2013 to Q4 2014, and testing from Q1 2015 to Q4 2016. The dataset was made from 23 years and the size is representative of many other tasks in NLP (Table 4).

The MDA section of an SEC Filing, despite only constituting a small part of the filing, is still very long. The average MDA section in our task is 13,267 tokens long using a BPE tokenizer (Sen-

---

nrich et al., 2016). As such, when a model was not able to encode all of the tokens, only the first part of the text was encoded.

4 Methodology

We deploy two frameworks to test different architectural methodologies on this task. The same data are provided to each of the frameworks. The first framework is a feature-based discriminative approach that uses a more traditional boosting-tree model and tests the different ways to encode the textual data. The second uses generative LLMs and prompting to output one of a fixed list of labels through a greedy search algorithm (App. A).

4.1 Task Description

The objective is to predict the credit rating, $\hat{R}_t$, at time t. The function can be represented as follows:

$$\hat{R}_t = f\big(T_{t-1}, T_{t-2}, \ldots, T_{t-p};$$
$$R_{t-1}, R_{t-2}, \ldots, R_{t-p};$$
$$N_{t-1}, N_{t-2}, \ldots, N_{t-p}\big)$$

Here, T_{t-i} represents the text data, R_{t-i} represents the historical credit rating data, N_{t-i} represents the numeric data - both fundamental and macroeconomic. i varies from 1 to p, with p indicating the number of past quarters considered (1 to 4 quarters in this study). Furthermore, f is the predictive function to convert the input data into an estimate. An ablation study is conducted to evaluate the impact of different data types on the prediction accuracy. In this study, the function f is tested under various configurations: using only text data T_{t-i}, using combinations of historical ratings R_{t-j}, and numeric data N_{t-k}. This approach helps to determine the relative importance of each type of data.

4.2 Boosting-Tree Baseline

To test the abilities of generative LLMs, it is necessary to benchmark the performance against a relatively well-understood and robust algorithm. We select XGBoost (Chen and Guestrin, 2016), a model that has been widely adopted in many domains (Talukder et al., 2023; Dong et al., 2023; Joshi et al., 2024). The supervised model takes as input the normalized fundamental, macroeconomic and text data, and outputs the most likely label. We describe other more complex neural network architectures that failed to learn this task in Appendix D. Due to the restricted dataset size, the models

outlined in the Appendix were unable to learn the task before overfitting the training dataset.

4.3 Text Encoders

To test and identify which of the traditional encoders performs best we trialled a series of standard methodologies.

The Loughran McDonald Lexicon (**LM**) (Loughran and McDonald, 2011) is widely recognized in finance. Given its prevalence, it is crucial to compare its effectiveness with more advanced methods. The lexicon classifies words into four domains: Positivity, Negativity, Litigiousness, and Uncertainty. However, the simple language modelling technique classifies phrases like "The debt increased last quarter" as neutral. The LM text representation in this work is the document word count from each sentiment, normalized by the maximum value in the training set.

Latent Dirichlet Allocation (**LDA**) is a widely-used topic modeling method that identifies latent topics within text (Blei et al., 2003). It operates by assuming that each document is a mixture of topics and that each topic is a distribution over words. Despite advancements in topic modeling, LDA remains a reliable baseline for evaluating newer models. In this paper, the features represent texts as probability distributions over 25 topics, with each dimension indicating the likelihood of the text belonging to a specific topic.

High-density Embedding Clusters (**HEC**) leverage the natural language understanding of LLMs but reduce the dimensionality of the input feature. HEC provides a good basis for topic modelling (Sia et al., 2020; Grootendorst, 2022). Sentence embeddings have also been used to discern domain type from text (Aharoni and Goldberg, 2020). Drinkall et al. 2022 extended this work to generate features from clusters of sentence embeddings in a COVID-19 caseload prediction task. For this task, each sentence of each filing in the training set was encoded into embeddings space using a *all-mpnet-base-v2* (Reimers and Gurevych, 2019), the dimensionality was then reduced using UMAP (McInnes et al., 2020), and the HDBSCAN clustering algorithm (Campello et al., 2013) was used for form 100 distinct clusters. An example of the cluster features is displayed in Figure 1. Then each filing in the train, validation and test set was split into sentences and then transformed into the embedding space described above. The overall text representation was the average of the representations of each sentence,

and the representation of each sentence was the probability distribution that the sentence belonged to each of the 100 clusters.

To understand the extent to which emotion scores play a significant role in forecasting the next credit rating. We used a DistilRoBERTa model (E_{DRoBERTa}) that had been fine-tuned on an emotion classification task (Hartmann, 2022). The SEC Filings are then chunked into 512 token sequences and classified according to the probability that that chunk can be associated with each emotion. The average across all chunks is taken as the final text representation of each filing.

We also trialled a pooled MP-NET representation (Song et al., 2020) by chunking the text into 512 token segments and averaging over the pooled representation of each chunk. In well-established benchmarks (Muennighoff et al., 2023), MP-NET embeddings have performed strongly for their size and provide a baseline comparison to the HEC features derived from the MP-NET model.

4.4 Generative Framework

Given recent advancements in generative LLMs, we evaluate whether these models can identify changes in a company's perceived risk and determine the best methodology for achieving high performance. This approach differs from other text encoding methods discussed earlier, as numerical data is converted into text format for the model to process using prompts. The prompts used in the following section are included in the Appendix E, and follow the best practice from existing literature (OpenAI, 2024; Lin et al., 2024). Sui et al. (2024) showed that contextual information about the tabular features enables a 0-shot framework to outperform 1-shot prompting methodology.

While LLMs perform very well in 0-shot settings (Kojima et al., 2023), there is significant evidence that shows that LLMs perform better in a k-shot setting (Clark et al., 2018); the ARC benchmark uses 25-shot prompts in the Eleuther AI evaluation harness (Gao et al., 2023). The problem with deploying a k-shot framework in this setting is that the SEC Filings are very long (13,267 tokens). Despite the increase in the context-window length of some newer LLMs, many new models are capped at 8192 tokens or below (Wang and Komatsuzaki, 2021; Jiang et al., 2023; Touvron et al., 2023; Grattafiori et al., 2024), and some other studies have shown performance deterioration as the input sequence increases (Li et al., 2024). Fitting several examples

of the task in the input sequence is impossible for many of the data samples, which means that k-shot performance is not reported for this task.

We tested several models[9] using the prompting structure laid out in Appendix E. The models provide a good representation of the current state-of-the-art (Chiang et al., 2024).

4.4.1 LoRA Adaptation

To adapt the LLMs we use LoRA (**Lo**w **R**ank **A**daptation) (Hu et al., 2021) fine-tuning. This technique involves optimizing the rank-decomposition matrices, A & B, of the change in model weights (ΔW), where W' are the new model weights and W are the pre-trained weights.

$$W' = W + \Delta W \qquad (1)$$
$$= W + BA \qquad (2)$$

The advantage is that it requires a lot less memory to fine-tune a model and in contrast to some parameter-efficient fine-tuning methods, adaptation can take place through the entire model stack.

5 Results

The results from the XGBoost baseline are outlined in Table 1. It is clear that there is some information in the text since almost all text encoding methods perform above chance. However, none of the individual text feature sets outperform the numeric

---

[9]*gpt-3.5-turbo-0125, gpt-4-turbo-2024-04-09* and *gpt-4o-2024-05-13, Llama-3 8B*

Data	Features	Av.	Quarters			
			1	2	3	4
N	$M + F + C$	**52.8**	**48.3**	**53.3**	**54.0**	**55.7**
	C	44.7	41.9	43.6	46.5	46.7
A	LM	50.6	46.8	51.2	52.1	52.4
	LDA	50.9	50.3	52.3	50.8	50.2
	HEC	<u>53.6</u>	<u>50.7</u>	**54.6**	54.1	**56.0**
	E_{DRoBERTa}	52.8	48.2	52.9	55.4	54.8
	MP-NET	51.0	46.8	52.5	<u>56.1</u>	48.4
T	LM	34.4	36.6	33.4	30.9	36.8
	LDA	34.8	36.6	35.0	30.3	**37.1**
	HEC	**38.1**	**39.8**	**38.0**	38.9	35.8
	E_{DRoBERTa}	36.0	36.8	35.8	36.2	35.2
	MP-NET	35.9	32.4	35.0	**39.9**	36.3

Table 1: The accuracy using the XGBoost model across different feature sets and text encoding methods. **N** refers to instances where only numeric information is used. **T** refers to text-based data types. **A** indicates all data types combined ($M + F + C + T$). **Bold** indicates the best results for each of the data configurations; <u>underline</u> indicates the best results across all configurations.

Model	Features	Average	Quarters				Feature Importances		
			1	2	3	4	**M**	**F**	**T**
XGBoost	**N**	52.8	48.3	53.3	54.0	55.7	0.659	0.341	-
	$\mathbf{A}_{\text{HEC}}$	**_53.9_**	50.7	_54.6_	_54.1_	56.0	0.117	0.188	0.695
	$\mathbf{T}_{\text{HEC}}$	38.1	39.8	38.0	38.9	35.8	-	-	-
	$\mathbf{N} + \mathbf{T}_{\text{GPT-4o Est.}}$	53.8	52.7	50.9	54.0	_57.3_	0.588	0.302	0.101
	$\mathbf{N}_{\text{XGB Est.}} + \mathbf{T}$	51.8	_53.7_	51.2	50.8	51.3	0.573		0.427
GPT-4o	**N**	31.4	33.7	31.6	31.1	29.3	-	-	-
	A	40.2	43.9	40.3	38.8	37.9	-	-	-
	T	**49.6**	**49.3**	**52.2**	**52.4**	**44.6**	-	-	-
	$\mathbf{T} + \mathbf{N}_{\text{XGB Est.}}$	32.3	33.9	30.8	36.7	27.7	-	-	-
XGBoost-**N** ∪ GPT-4o-**T**		69.9	70.5	73.2	71.5	64.3	-	-	-

Table 2: Accuracy across different model and data configurations. The notation is consistent to Table 1. Feature importances for $\mathbf{M}, \mathbf{F}, \mathbf{T}$ are impurity scores averaged across lags. $\mathbf{N}_{\text{XGB Est.}}$ and $\mathbf{T}_{\text{GPT-4o Est.}}$ represent the subscript model's estimate and implied internal probability of that estimate using the features represented by the bold letter, both are the probability and estimate are used as features.

Data	Model	Av.	Quarters			
			1	2	3	4
N	Llama	32.3	35.1	**35.8**	29.0	29.3
	Llama-LoRA	**35.5**	35.4	34.2	**37.8**	34.7
	GPT-3.5	32.6	32.9	33.4	31.9	32.3
	GPT-4	34.1	34.2	32.9	33.0	**36.3**
	GPT-4o	31.4	33.7	31.6	31.1	29.3
A	Llama	35.5	35.4	35.8	37.0	33.9
	Llama-LoRA	37.5	36.4	37.0	37.6	38.8
	GPT-3.5	**44.5**	**49.0**	42.4	**46.0**	40.6
	GPT-4	38.3	39.8	36.1	38.0	39.3
	GPT-4o	40.2	43.9	40.3	38.8	37.9
T	Llama	35.6	35.4	36.1	37.0	33.9
	Llama-LoRA	37.0	38.1	36.8	37.1	35.9
	GPT-3.5	46.4	47.3	47.5	45.5	45.2
	GPT-4	48.5	47.8	48.5	50.3	**_48.1_**
	GPT-4o	**49.6**	**49.3**	**52.2**	**52.4**	44.6

Table 3: Accuracy using the generative models. The notation is consistent to Table 1. All models are tested in 0-shot besides Llama-3 8B which is fine-tuned using LoRA.

baselines, indicating that fundamental and macroeconomic variables are more critical for prediction. Combining features yields a performance boost, particularly when HEC features are integrated with numerical data.

Table 3 highlights intriguing behavior in generative models. With a zero-shot prompt, performance using only numerical data is near random. Interestingly, GPT-class models perform better using text alone than with all data types, suggesting that numerical information may hinder their predictive accuracy. GPT-3.5, despite being older, achieves the best performance with all features. It also appears that LoRA enables better relative performance on numerical data - Llama-3 8B LoRA is the best-performing model on entirely numerical informa-

tion and is the only model with no performance degradation when all features are considered as opposed to just text. Overall, generative models excel at decoding text data for this task.

Table 2 takes the best-performing text features from the XGBoost framework, HEC, and provides a comparison to the best-performing generative model, GPT-4o. Interestingly, GPT-4o utilises the text alone much better than any of the encoder-based methods, but when all feature-types are considered the XGBoost-HEC configuration is the best-performing methodology.

In addition, the models pick up on different signals. The final row of Table 2 shows that the proportion of samples where at least one of the XGBoost-**N** and GPT-4o-**T** is correct (69.87) is significantly higher than any of the individual models. As a result, we provide comparisons where the estimate and class probability of the XGBoost-**N** and GPT-4o-**T** are included in the prompt or feature set. Both the estimate and the probability that each model assigns to the estimate are used in the prompt or as features. The combination methods all underperform the XGBoost-$\mathbf{A}_{\text{HEC}}$ configuration, but this performance gap provides an opportunity for future research on ensembling methods.

6 Interpretability

One of the disadvantages of generative LLMs is that, to a large degree, they are black boxes. While some work has successfully used attention weights and internal model states to analyse generated prompts (Serrano and Smith, 2019; Wang et al., 2022), this is not a mature research area. Much of the mechanistic interpretability literature has

focused on toy models (Elhage et al., 2021), and while there has been some progress made on extracting features from larger models using sparse autoencoders (Templeton et al., 2024), the field is still very far from completely solving interpretability within LLMs. In the absence of a complete solution, it is worth acknowledging the interpretable features that traditional methods use. Regulation in major economies increasingly emphasizes explainability alongside performance (European Commission, 2020; US Congress, 2022; UK Secretary of State for DCMS, 2022). The XGBoost-$\mathbf{A}_{\text{HEC}}$ framework not only achieves the best performance among the models in this paper but also enables users to interpret its decisions. This section provides an example of how we can use this framework to understand the reasons behind decisions.

The most obvious advantage of a feature-based system is that important features can be identified. Table 2 provides an example of feature importance that can be used to infer the modality preference of model configurations. It is possible to conduct even more granular feature analysis by looking at the contribution of individual features. Figure 2 shows the partial dependence plot of some of the individual text features on the "Up" and "Down" classes. From the plot in Figure 2a we can infer that as ratings are discussed more in a company's filing, there is a reduced chance of the the credit rating being upgraded. We can also infer from Figure 2b that as companies talk about receivables - the money owed to the company - there is a reduced chance of the credit rating being downgraded. Both provide valuable insights and are examples of how a traditional feature-based methodology can be leveraged for increased interpretability.

7   Conclusion

The paper shows that while LLMs are good at encoding textual data and inferring signals that traditional methods cannot pick up on when combined with numerical data in the prompt there is performance deterioration. The other advantage is that traditional methods offer increased interpretability and a better understanding of the mechanisms behind certain predictions. In addition, traditional approaches don't suffer to the same extent from complications associated with training data contamination and memorization (Ozdayi et al., 2023; Lu et al., 2024) since the models used for the traditional features are much smaller than the generative
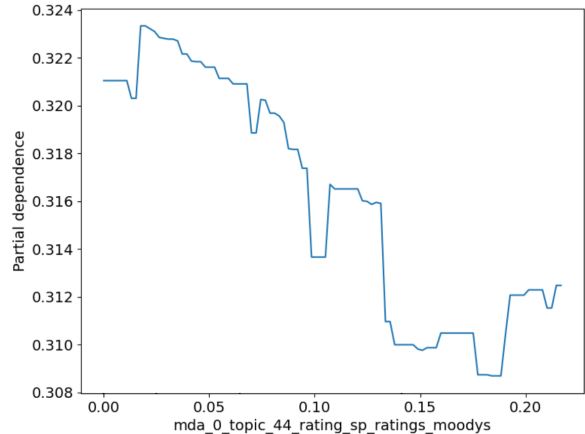
(a) PDP of "rating_sp_ratings_moodys" cluster & "Up" class.
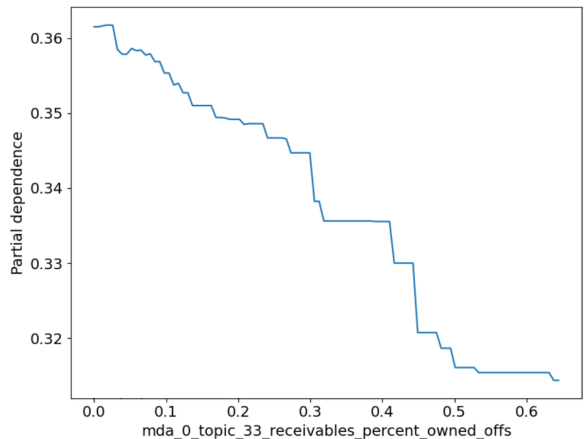
(b) PDP of "receivables_percent_owned_offs" cluster & "Down" class.

Figure 2: Partial Dependence Plots (PDP) of text-based features against different target classes.

models and memorisation in LLMs exhibits scaling law behaviour. While it is not impossible that the text data used in this paper was included in the training of the studied LLMs, any potential influence would likely have a greater impact on generative models, thereby reinforcing the findings of this paper.

There has been some work jointly encoding text using generative LLMs with time-series information (Liu et al., 2024), but more work needs to be done to determine the best methodology for combining long text sequences with numerical information while utilizing the benefits of generative LLM natural language understanding. This paper shows that combining multimodal information within the prompt is not sufficient.

125

8 Limitations

The task above uses a balanced dataset, which is good for testing the different methodologies' ability to discern the signals that are predictive of a rise or fall in credit ratings, but is poor for assessing how good the models would be in a real-world context where almost all of the ratings stay the same. Despite the data being taken from across all US equities over a 23-year time period, the balanced dataset is relatively small, with only 3441 samples for the Lag 1 configuration and 2142 samples for the Lag 4 configuration. There are plenty of prominent datasets that are smaller, but the size reduces the scope for complex and specialized models to be deployed on this task in favour of more robust, simple models.

Another limitation is that the text used in this paper is produced by the companies themselves, who the goal of conveying a positive viewpoint to investors. More objective publication venues may produce different insights about the future direction of a company.

We also assume that the credit rating methodology remains the same between the train and test sets. This is an assumption that is made by the rest of the literature, and our training set is spread over an 18 year period, however it does not rule out the possibility that the results are only valid over the time period that was tested. Due to the size of the dataset we were restricted from using a masked temporal cross-validation evaluation framework, which would have left insufficient data for training for some years.

The LoRA fine-tuning methodology outlined in this paper is a parameter-efficient technique and has been shown to be competitive in a variety of settings (Hu et al., 2021), but can be outperformed in some tasks by full fine-tuning and other adapter-based methods (Xu et al., 2023). We compared the performance of the LoRA implementation in this paper to that of QLoRA (Dettmers et al., 2023), which produced marginally worse results. However, it is possible that other fine-tuning techniques would have produced better results. Further to this, it is possible that if a better model than Llama-3 8B had been fine-tuned we would have seen even better results from the generative LLMs. The computational constraints placed on us are not dissimilar to those that other researchers face, which makes the results in this paper valid while perhaps not exhaustive.

References

Abien Fred Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU). *Preprint*, arXiv:1803.08375.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Wuyue An, Lin Wang, and Yu-Rong Zeng. 2023. Text-based Soybean Futures Price Forecasting: A Two-stage Deep Learning Approach. *Journal of Forecasting*, 42(2):312–330.

Yalanati Ayyappa, B. Vinay Kumar, Sudha-batthula Padma Priya, et al. 2023. Forecasting Equity Prices using LSTM and BERT with Sentiment Analysis. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, pages 643–648.

Golnoosh Babaei and Paolo Giudici. 2024. GPT Classifications, with Application to Credit Lending. *Machine Learning with Applications*, 16:100534.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Luca Barbaglia, Sergio Consoli, and Sebastiano Manzan. 2023. Forecasting with Economic News. *Journal of Business & Economic Statistics*, 41(3):708–719.

Suzana Baresa, Sinisa Bogdan, and Sasa Ivanovic. 2012. Role, Interests and Critics of Credit Rating Agencies. *UTMS Journal of Economics*, 3(1):71–82.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *Preprint*, arXiv:2004.05150.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Romain Boulland, François Degeorge, and Edith Ginglinger. 2016. News Dissemination and Investor Attention*. *Review of Finance*, 21(2):761–791.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.

Defu Cao, Furong Jia, Sercan O Arik, et al. 2023. TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting. *Preprint*, arXiv:2310.04948.

Yupeng Cao, Zhi Chen, Qingyun Pei, et al. 2024. RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data. *Preprint*, arXiv:2404.07452.

Kenneth Carling, Tor Jacobson, Jesper Lindé, et al. 2007. Corporate Credit Risk Modeling and the Macroeconomy. *Journal of Banking & Finance*, 31(3).

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. LLM4TS: Two-Stage Fine-Tuning for Time-Series Forecasting with Pre-Trained LLMs. *Preprint*, arXiv:2308.08469.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, et al. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Shouyuan Chen, Sherman Wong, Liangjian Chen, et al. 2023. Extending Context Window of Large Language Models via Positional Interpolation. *Preprint*, arXiv:2306.15595.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Yuh-Jen Chen and Yuh-Min Chen. 2022. Forecasting Corporate Credit Ratings using Big Data from Social Media. *Expert Syst. Appl.*, 207(C).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *Preprint*, arXiv:2403.04132.

Hyungjin Cho, Seung-Youb Han, Seungbin Oh, et al. 2020. Optimistic Credit Rating and its Influence on Corporate Decisions: Evidence from Korea. *Asia-Pacific Journal of Accounting & Economics*, 27(5):612–629.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *Preprint*, arXiv:1412.3555.

Peter Clark, Isaac Cowhey, Oren Etzioni, et al. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianhua Dong, Wenzhi Zeng, Lifeng Wu, et al. 2023. Enhancing Short-Term Forecasting of Daily Precipitation Using Numerical Weather Prediction Bias Correcting with XGBoost in Different Regions of China. *Engineering Applications of Artificial Intelligence*, 117:105579.

Felix Drinkall, Eghbal Rahimikia, Janet Pierrehumbert, and Stefan Zohren. 2024. Time Machine GPT. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3281–3292, Mexico City, Mexico. Association for Computational Linguistics.

Felix Drinkall, Stefan Zohren, and Janet Pierrehumbert. 2022. Forecasting COVID-19 Caseloads Using Unsupervised Embedding Clusters of Social Media Posts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. 2021. A Mathematical Framework for Transformer Circuits.

European Commission. 2020. White Paper: On Artificial Intelligence- A European Approach to Excellence and Trust. Accessed: 09.07.24.

Marzieh Fadaee and Christof Monz. 2020. The Unreasonable Volatility of Neural Machine Translation Models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Wenying Fei, Jing Gu, Yang Yang, et al. 2015. Credit Risk Evaluation Based on Social Media. *Procedia Computer Science*, 55:725–731. 3rd International Conference on Information Technology and Quantitative Management, ITQM.

Koresh Galil, Ami Hauptman, and Rosit Levy Rosenboim. 2023. Prediction of Corporate Credit Ratings with Machine Learning: Simple Interpretative Models. *Finance Research Letters*, 58:104648.

Leo Gao, Jonathan Tow, Baber Abbasi, et al. 2023. A Framework for Few-Shot Language Model Evaluation.

David Gillmor. 2015. Standard & Poor's Rating Process. Accessed: 01.11.23.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.

Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. https://tinyurl.com/mtmh4btv. Accessed: 09.07.24.

Dan Hendrycks, Collin Burns, Steven Basart, et al. 2021. Measuring Massive Multitask Language Understanding. *Preprint*, arXiv:2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065. Association for Computational Linguistics.

Mikella Hurley and Julius Adebayo. 2016. Credit Scoring in the Era of Big Data. *Yale JL & Tech.*, 18:148.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Karen Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*.

Anushka Joshi, Balasubramanian Raman, C Krishna Mohan, et al. 2024. Application of a New Machine Learning Model to Improve Earthquake Ground Motion Predictions. *Natural Hazards*, 120(1):729–753.

Jean Kaddour, Joshua Harris, Maximilian Mozes, et al. 2023. Challenges and Applications of Large Language Models. *Preprint*, arXiv:2307.10169.

Eleni Kalamara, Arthur Turrell, Chris Redl, et al. 2022. Making Text Count: Economic Forecasting using Newspaper Text.

Nont Kanungsukkasem and Teerapong Leelanupab. 2019. Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction. *IEEE Access*, 7:71645–71664.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, et al. 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*, pages 272–280.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, et al. 2023. Large Language Models are Zero-Shot Reasoners. *Preprint*, arXiv:2205.11916.

Geon Lee, Wenchao Yu, Wei Cheng, et al. 2023. MoAT: Multi-Modal Augmented Time Series Forecasting.

Jingyuan Li, Caosen Xu, Bing Feng, et al. 2023. Credit Risk Prediction Model for Listed Companies Based on CNN-LSTM and Attention Mechanism.

Tianle Li, Ge Zhang, Quy Duc Do, et al. 2024. Long-context LLMs Struggle with Long In-context Learning. *Preprint*, arXiv:2404.02060.

Xiaodong Li, Haoran Xie, Li Chen, et al. 2014. News Impact on Stock Price Return via Sentiment Analysis. *Knowledge-Based Systems*, 69:14–23.

Charalampos M Liapis and Sotiris Kotsiantis. 2023. Temporal Convolutional Networks and BERT-Based Multi-Label Emotion Analysis for Financial Forecasting. *Information*, 14(11):596.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, et al. 2024. Graph-enhanced Large Language Models in Asynchronous Plan Reasoning. *Preprint*, arXiv:2402.02805.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Preprint*, arXiv:2109.07958.

Xu Liu, Junfeng Hu, Yuan Li, et al. 2024. UniTime: A Language-Empowered Unified Model for Cross-Domain Time Series Forecasting. *Preprint*, arXiv:2310.09751.

Mark Lokanan, Vincent Tran, and Nam Hoai Vuong. 2019. Detecting Anomalies in Financial Statements Using Machine Learning Algorithm: The Case of Vietnamese Listed Firms. *Asian Journal of Accounting Research*, 4(2):181–201.

Isabelle Lorge and Janet B. Pierrehumbert. 2023. Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models. In *Proceedings of the 6th BlackboxNLP Workshop*.

Tim Loughran and Bill McDonald. 2011. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Wiley Online Library.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling Laws for Fact Memorization of Large Language Models. *Preprint*, arXiv:2406.15720.

Feng Mai, Shaonan Tian, et al. 2019. Deep Learning Models for Bankruptcy Prediction Using Textual Disclosures.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Preprint*, arXiv:1802.03426.

Saif M. Mohammad. 2020. Practical and Ethical Considerations in the Effective use of Emotion and Sentiment Lexicons. *Preprint*, arXiv:2011.03492.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Nora Muñoz-Izquierdo, María Jesús Segovia-Vargas, María-del-Mar Camacho-Miñano, et al. 2022. Machine Learning in Corporate Credit Rating Assessment Using the Expanded Audit Report. volume 111, pages 4183–4215. Springer.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, et al. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *Preprint*, arXiv:2406.11903.

OpenAI. 2024. Best Practices for Prompt Engineering with the OpenAI API. Online. Accessed: 10.07.24.

Mustafa Ozdayi, Charith Peris, Jack FitzGerald, et al. 2023. Controlling the Extraction of Memorized Data from Large Language Models via Prompt-Tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1512–1521.

Mustafa Pamuk and Matthias Schumann. 2023. Opening a New Era with Machine Learning in Financial Services? Forecasting Corporate Credit Ratings Based on Annual Financial Statements. volume 11.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, et al. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *Preprint*, arXiv:2112.11446.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, et al. 2017. Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Khaladdin Rzayev, Gbenga Ibikunle, and Tom Steffen. 2023. The Market Quality Implications of Speed in Cross-Platform Trading: Evidence from Frankfurt-London Microwave Networks. volume 66, page 100853.

Oscar Sainz, Jon Campos, Iker García-Ferrero, et al. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, et al. 2020. Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pretraining for Language Understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Minchae Song and Kyung-shik Shin. 2019. Forecasting Economic Indicators Using a Consumer Sentiment Index: Survey-based Versus Text-based Data. volume 38, pages 504–518.

Yejun Soun, Jaemin Yoo, Minyong Cho, et al. 2022. Accurate Stock Movement Prediction with Self-supervised Learning from Sparse Noisy Tweets. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1691–1700. IEEE.

S&P-Global. 2016. S&P Global Ratings Definitions. RatingsDirect. Accessed: 10.07.24.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. 2023. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Preprint*, arXiv:2206.04615.

Matthew Stevenson, Christophe Mues, and Cristián Bravo. 2021. The Value of Text for Small Business Default Prediction: A Deep Learning Approach. *European Journal of Operational Research*, 295(2):758–771.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large Language Models are Inconsistent and Biased Evaluators. *Preprint*, arXiv:2405.01724.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, et al. 2024. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Md Alamin Talukder, Khondokar Fida Hasan, Md Manowarul Islam, et al. 2023. A Dependable Hybrid Machine Learning Model for Network Intrusion Detection. *Journal of Information Security and Applications*, 72:103405.

Mahsa Tavakoli, Rohitash Chandra, Fengrui Tian, et al. 2023. Multi-Modal Deep Learning for Credit Rating Prediction Using Text and Numerical Data Streams. *Preprint*, arXiv:2304.10740.

Yi Tay, Mostafa Dehghani, Dara Bahri, et al. 2022. Efficient Transformers: A Survey. *ACM Comput. Surv.*

Mark P. Taylor, Zigan Wang, and Qi Xu. 2021. The Real Effects of Exchange Rate Risk on Corporate Investment: International Evidence. volume 117, page 102432.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, et al. 2023. Language Models are not Naysayers: an Analysis of Language Models on Negation Benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*.

UK Secretary of State for DCMS. 2022. AI Regulation Policy Paper. Accessed: 09.07.24.

US Congress. 2022. Algorithmic Accountability Act of 2022. Accessed: 09.07.24.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. Accessed: 09.07.24.

Kevin Wang, Alexandre Variengien, Arthur Conmy, et al. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *Preprint*, arXiv:2211.00593.

Lu Wang. 2022. Imbalanced Credit Risk Prediction Based on SMOTE and Multi-Kernel FCM Improved by Particle Swarm Optimization.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yiren Wang, Dominic Seyler, Shubhra Kanti Karmaker Santu, et al. 2017. A Study of Feature Construction for Text-based Forecasting of Time Series Variables. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2347–2350.

Qianqian Xie, Weiguang Han, Xiao Zhang, et al. 2024. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *Preprint*, arXiv:2312.12148.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, et al. 2024. Benchmarking Benchmark Leakage in Large Language Models. *Preprint*, arXiv:2404.18824.

Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Hao Xue and Flora D. Salim. 2024. PromptCast: A New Prompt-Based Learning Paradigm for Time Series Forecasting . *IEEE Transactions on Knowledge & Data Engineering*, 36(11):6851–6864.

Linyi Yang, Tin Lok James Ng, Barry Smyth, et al. 2020. HTML: Hierarchical Transformer-based Multi-task Learning for Volatility Prediction. In *Proceedings of The Web Conference*, pages 441–451.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for Temporal Data - A Study on Explainable Financial Time Series Forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753.

Xi Zhang, Yunjia Zhang, Senzhang Wang, et al. 2018. Improving Stock Market Prediction via Heterogeneous Information Fusion. volume 143. Elsevier BV.

Zixue Zhao, Tianxiang Cui, Shusheng Ding, et al. 2024. Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction. *Mathematics*.

Tian Zhou, Peisong Niu, Xue Wang, et al. 2024. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Barret Zoph, Colin Raffel, Dale Schuurmans, et al. 2022. Emergent Abilities of Large Language Models. *TMLR*.

A Greedy Decoding

Generative models can produce unpredictable outputs (Fadaee and Monz, 2020; Stureborg et al., 2024), which necessitates the use of constrained generation when an LLM forms part of a larger architecture. For the purposes of this work, we use a greedy search to infer the probability of one of the following labels appearing next in the sequence: "up", "down" or "same".

B Dataset Size

The size of each of the dataset splits are outlined in Table 4.

# Quarters	Train	Val	Test
1	2,642	389	410
2	1,748	374	377
3	1,595	351	376
4	1,445	325	372

Table 4: Dataset sizes

C Fundamental Data

The fundamental variables considered are outlined in Table 5.

Variable	Type	Description
niq	Float	Net Income (Loss)
ltq	Float	Liabilities - Total
piq	Float	Pretax Income
atq	Float	Assets - Total
ggroup	Char	GIC Groups
gind	Char	GIC Industries
gsector	Char	GIC Sectors
gsubind	Char	GIC Sub-Industries

Table 5: Description of Variables

D Neural Network Implementations

We also tested some more complex neural network (NN) approaches, which had underwhelming results. The **Hier**archical **C**redit **R**ating (HierCR) model is a framework that models the filings hierarchically. The challenge with using LLMs to encode the filings is the limited context window of encoder-based LLMs. There have been many different solutions to this problem, including sparse attention mechanisms (Beltagy et al., 2020), chunking (Sawhney et al., 2020), and feature-based extraction like the methods above (Loughran and McDonald, 2011; Drinkall et al., 2022). Our NN solution to this problem is to split the filing up into sentences and pass the sentence embeddings through an *all-mpnet-base-v2* encoder to produce embeddings for the textual data. The text encoder replicates the structure in (Sawhney et al., 2020), the only material difference is that filing sentences substitute the social media posts in the first layer of the text-encoder. The text, macro and fundamental vectors across the previous quarter(s) are combined using a GRU layer (Chung et al., 2014), the outputs are then passed through an attention layer to create a representation for each data type. These representations are combined using a bilinear transformation, which is passed through 3 linear layers followed by a ReLU (Agarap, 2019) activation function. Dropout is applied in the final linear-layer classification module.

Model	Av.	Quarter			
		1	2	3	4
SA-LF	40.62	39.87	40.05	41.17	41.39
SA-EF	43.41	41.24	40.97	45.31	46.11
HierCR	35.02	33.58	32.89	35.62	37.98

Table 6: Accuracy for more complex NN approaches using all data types (M+F+T).

The other two architectures are **S**hared **A**ttention **L**ate **F**usion (SA-LF) and **S**hared **A**ttention **E**arly **F**usion (SA-EF). Both architectures only consider the first 512 tokens of each filing. The difference between the two architectures is when the attention layer is applied. For the SA-EF the model attends to all feature types together, whereas the SA-LF only combines the representations after the attention layer is applied to the individual data types. The final representation is the passed through the same linear-layer classification module as the HierCR. For all of the architectures above, we trained the model for 200 epochs with a patience value of 20 epochs.

The results from these models are poor in comparison to the more simple XGBoost models. This could be due to the size of the dataset, which does not provide the model enough data to train on without overfitting the training data. Complex models with many parameters require more data to fit properly. Given that the dataset is the largest balanced and complete dataset possible to make using US data, and that the size of the dataset considered in this paper is representative of a large number of other tasks, the results from this paper represent a

significant contribution for dealing with problems of this nature.

E Prompts

To encode the numerical and textual information into text form, we used the prompting structure outlined below. When the ablation study was carried out the prompt and data included was adjusted accordingly.

System: *You are trying to work out whether a company's credit rating is likely to go up, down, or stay the same given its recent credit ratings. Predict the likely movement in a company's credit rating for the next quarter, using historical credit ratings, quantitative financial data and macroeconomic data. The numeric data has been normalized and appears in order with the most recent first.*
Credit Rating Explanation:
Credit ratings use the following scale, in order of increasing risk: 'AAA', 'AA+', 'AA', 'AA-', 'A+', 'A', 'A-', 'BBB+', 'BBB', 'BBB-', 'BB+', 'BB', 'BB-', 'B+', 'B', 'B-', 'CCC', 'CCC-', 'CC', 'C', 'SD'
Fundamental Financial Indicators Defined:
...
Macroeconomic Variables Defined:
...
User:
Your task is to classify the company into one of the following classes: "down", "same", "up". "down" means that you think the credit rating will go down in the next quarter, meaning the company is perceived as more risky. "same" means that you think the credit rating will stay the same in the next quarter. "up" means that you think the credit rating will go up in the next quarter, meaning the company is perceived as less risky. Please respond with a single label that you think fits the company best.
Classify the following numerical data:"""

E.1 Credit Rating Ranking

One potential problem with the prompt outlined in Appendix E is that the LLM may find it hard to correctly understand the ranking structure of credit ratings, which would limit the ability of an LLM to perform well on this task. To probe the LLMs ability to understand the relative rank of credit ratings we created the following prompt:

"""Two credit ratings will be given, the task is to determine which is higher on the following scale, which is ordered in descending order:

'AAA', 'AA+', 'AA', 'AA-', 'A+', 'A', 'A-', 'BBB+', 'BBB', 'BBB-', 'BB+', 'BB', 'BB-', 'B+', 'B', 'B-', 'CCC', 'CCC-', 'CC', 'C', 'SD'.

Please answer with the higher rating e.g. AAA vs. SD Answer: AAA.
«rating_X» vs. «rating_Y» Answer:"""

The performance on this task across all rating combinations when prompting GPT-4o was 99.52%. The only mistake was between C and CC. This high performance displays a very good understanding of the credit rating scale and justifies the setup of our prompt.

F S&P Credit Rating Definitions

S&P's definitions for each of the credit rating categories are outlined in Table 7.

Category	Definition
AAA	An obligation rated 'AAA' has the highest rating assigned by S&P Global Ratings. The obligor's capacity to meet its financial commitment on the obligation is extremely strong.
AA	An obligation rated 'AA' differs from the highest-rated obligations only to a small degree. The obligor's capacity to meet its financial commitment on the obligation is very strong.
A	An obligation rated 'A' is somewhat more susceptible to the adverse effects of changes in circumstances and economic conditions than obligations in higher-rated categories. However, the obligor's capacity to meet its financial commitment on the obligation is still strong.
BBB	An obligation rated 'BBB' exhibits adequate protection parameters. However, adverse conditions or changing circumstances are likely to lead to a weakened capacity of the obligor to meet its financial commitment on the obligation.
BB; B; CCC; CC; and C	Obligations rated 'BB', 'B', 'CCC', 'CC', and 'C' are regarded as having significant speculative characteristics. 'BB' indicates the least degree of speculation and 'C' the highest. While such obligations will likely have some quality and protective characteristics, these may be outweighed by large uncertainties or major exposures to adverse conditions.
BB	An obligation rated 'BB' is less vulnerable to nonpayment than other speculative issues. However, it faces major uncertainties or exposure to adverse business, financial, or economic conditions which could lead to the obligor's inadequate capacity to meet its financial commitment on the obligation.
B	An obligation rated 'B' is more vulnerable to nonpayment than obligations rated 'BB', but the obligor currently has the capacity to meet its financial commitment on the obligation. Adverse business, financial, or economic conditions will likely impair the obligor's capacity or willingness to meet its financial commitment on the obligation.
CCC	An obligation rated 'CCC' is currently vulnerable to nonpayment, and is dependent upon favorable business, financial, and economic conditions for the obligor to meet its financial commitment on the obligation. In the event of adverse business, financial, or economic conditions, the obligor is not likely to have the capacity to meet its financial commitment on the obligation.
CC	An obligation rated 'CC' is currently highly vulnerable to nonpayment. The 'CC' rating is used when a default has not yet occurred, but S&P Global Ratings expects default to be a virtual certainty, regardless of the anticipated time to default.
C	An obligation rated 'C' is currently highly vulnerable to nonpayment, and the obligation is expected to have lower relative seniority or lower ultimate recovery compared to obligations that are rated higher.
SD	An obligation rated 'SD' is in default or in breach of an imputed promise. For non-hybrid capital instruments, the 'SD' rating category is used when payments on an obligation are not made on the date due, unless S&P Global Ratings believes that such payments will be made within five business days in the absence of a stated grace period or within the earlier of the stated grace period or 30 calendar days.
NR	This indicates that no rating has been requested, or that there is insufficient information on which to base a rating, or that S&P Global Ratings does not rate a particular obligation as a matter of policy.

Table 7: S&P Global Ratings Definitions (S&P-Global, 2016)