# "Leveraging Large Language Models for Real-time Financial Crisis Identification "

Koveshnikov, Viktor

## ABSTRACT

This thesis examines the use of large language models (LLMs) for real-time financial crisis identification using textual data from OECD reports. The study explores the possibility of applying state-of-the-art models to textual data to solve some of the limitations which are attributed to traditional numerical models. Meta's LLAMA-2 model was prompted to analyse the economic reports and decide whether a particular report described a period of a financial crisis. The model achieved an accuracy rate of over 91% and a ROC-AUC score of 0.836. Results demonstrate that LLMs can effectively identify financial crises, offering a valuable complement to numerical models. This research enhances the literature on NLP applications in economics and proposes a novel component, which could be added to early warning systems employed by central banks and policymakers.

## CITE THIS VERSION

UCLouvain

Economics School of Louvain - ESL

# Leveraging Large Language Models for Real-time Financial Crisis Identification

Author: Viktor Koveshnikov

Thesis Director: Prof. Francesca Monti

Thesis Reader: Prof. Christophe Corro

Academic Year: 2023-2024

Master in Economics – 120 credits – Econometrics Focus

Erasmus Mundus Joint Master Degree in Models and Methods of Quantitative Economics

Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve

# Contents

# Introduction

Financial crises disrupt societies. Not only do they result in a significant deceleration of economic growth, but they also lead to undesirable political and social consequences. Real prices on housing and stock indices fall dramatically, unemployment rates reach all-time highs and government debts suffer from the reduction of tax revenues and an abundance of countercyclical fiscal measures (Reinhart & Rogoff, 2009).

After the Global Financial Crisis of 2008-2009 central banks all over the world turned to macroprudential policies in order not to allow such a financial catastrophe to occur again. Macroprudential strategies are intended to mitigate the risks posed by financial instability. Numerous measures may be implemented as part of those strategies, including limiting debt-to-value ratios, setting high liquidity coverage ratios, and putting in place tighter levels of lending standards (Hanson et al., 2010; Shin, 2011). However, undertaking these measures is a matter of whether a central bank believes there is enough risk accumulated in the economy. Identifying the right moment to act is key for the macroprudential measures to work effectively.

Central banks have developed various models with the aim of constructing an accurate indicator that would give them warning signals whenever the risks become too high for the system as a whole. With the onset of Big Data, machine learning models, which extensively use macroeconomic and financial indicators, are emerging as widely used tools for this cause and they have demonstrated some sufficient results. The issue is that central banks need the warning signals in real-time, while some quantitative data for standard machine learning models may take months or even years to get published and updated. Indeed, central banks also monitor the day-to-day conditions on the stock, currency, and real estate markets, however, that may not provide them with the full picture.

The approach which is examined in the current paper concerns the analysis of textual data using natural language processing (NLP) and large language models (later – LLMs) with the aim of developing an early warning system for central banks and governments. This study is to demonstrate that qualitative textual data possesses insights different from those in quantitative data, it is able to provide more interpretable results and, therefore, its analysis may complement the numerical models currently being used by the majority of central banks. The research on the application of NLP to identify financial crises is rather limited, and this study will enrich the literature on the topic as well as provide models for practical use.

Given this context, the research question of the paper is:

**Are large language models able to identify financial crises based on the texts of economic reports?**

The results of the paper were obtained with the LLAMA-2 model (Touvron et al., 2023) and prompt engineering. In the classification task of identifying financial crises within a heavily unbalanced dataset of 1104 reports, the LLM accurately identified crises in 110 cases and correctly recognized stability in 902 instances, resulting in an accuracy rate of over 91% and a ROC-AUC score of 0.836, showing comparable performance to traditional machine learning models. Analysis of errors and key words as well as sentiment analysis of the model's explanations revealed that the model effectively linked economic indicators to broader contexts, however, it occasionally misinterpreted recovery signals as economic stability or confused severe recessions with financial crises due to the lack of a specific methodology.

Section 1 provides a comprehensive overview of papers on identification and prediction of financial crises as well as on the use of NLP in the economics literature. Section 2 is a brief description of the theoretical framework, which is followed by a detailed description of the data and the empirical strategy in Section 3. Section 4 provides the main results of the model and their interpretation. Section 5 presents the discussion and limitations of the research.

# Literature Review

One of the only reliable ways to identify a crisis until recently has involved consulting experts and trusting their judgement. This approach has resulted in comprehensive papers and corresponding databases, which are in the best-case scenario updated once in several years (Laeven & Valencia, 2020; Reinhart & Rogoff, 2009; Romer & Romer, 2017). In addition to slow updates, these databases are constructed retrospectively and rely on the available numerical data. However, it is crucial to have the warning system model functioning in real-time for macroprudential policies to mitigate risks and prevent potential crises effectively.

Models designed to alert authorities of increased systemic risk and a potential financial crisis have been in development since the late 1990s. Early models were constructed in

such a way that their outcomes were merely related to a single financial indicator, usually linearly (Eichengreen & Rose, 1998; Kaminsky et al., 1998; Sachs et al., 1996; Sarkar & Patel, 1998). Over time, these models have undergone refinement and enhancement: the signalling approach of Kaminsky et al. (1998) was modified by Knedlik and Von Schweinitz (2012), while Duca and Peltonen (2013) have worked on increasing the quality of logit/probit models. These models, however, have been surpassed by more sophisticated ones, capable of capturing non-linear relationships and incorporating interaction effects among a number of indicators. Sevim et al. (2014) argue that artificial neural networks (ANNs) consistently outperform traditional regression models, while Holopainen and Sarlin (2015) have demonstrated that decision trees, k-nearest-neighbours, support vector machines (SVM), and ANNs are all sufficient algorithms for the purpose of constructing an early warning system. Tanaka et al. (2016) and Alessi and Detken (2018) have looked at ensemble algorithms, more specifically random forests, and they as well as all aforementioned algorithms have resulted in better quality metrics than the signalling approach and the logit models. Beutel et al. (2019) show that, nevertheless, logit models do perform rather well on out-of-sample predictions and are not likely to overfit, unlike many other modern machine learning models.

In terms of most recent research, Samitas et al. (2020) have obtained an outstanding accuracy of 98.8% on their dataset of financial data, using structured financial networks and machine learning. Tölö (2020) has applied recurrent neural networks and achieved ROC-AUC of 0.75 on out-of-sample forecasts on the Jordá-Schularick-Taylor Macrohistory database (Jordà et al., 2017). On the same dataset, Bluwstein et al. (2020) achieve ROC-AUC scores between 0.82 and 0.87, using various machine learning algorithms (logit, SVM, random forest, extreme trees, etc.). Fouliard et al. (2020) try to combine numerous models and propose a meta-statistical approach of model aggregation with similar results in terms of the ROC-AUC score, depending on the country of investigation.

All of the models described above, however, are numerical models, and they face some limitations which are described in the literature in detail, specifically in Chen et al. (2023). A majority of them are based on indexes that describe whether there are any market disruptions in a given country at a given moment or not (Brave & Butters, 2012; Drehmann & Juselius, 2014; Lee et al., 2020). What these indexes capture, and that consequently reflects on the machine learning models which use them as predictors, is only the beginning stage of a financial crisis, while modern financial crises may last up to several years. This particular issue leaves researchers experimenting with models which would not be triggered by certain figures on financial markets but rather, in addition to that, would also consider "the big picture". Besides, numerical models are also dependent on the methodology which has been used to calculate particular variables and which may vary from year to year, conditional on

the government's regulations. Finally, numerical models are updated slowly, again due to the nature of their input data.

One of the solutions, aimed at mitigating all the downsides of modern numerical models, is the use of textual data, natural language processing (NLP), and machine learning models which take vectorized textual data as input.

Texts have descriptive properties which numbers do not: "the unemployment rate of 10%" and "the unemployment rate reached an all-time high, millions of people cannot find a job" are two significantly different expressions, although they may refer to the same situation. Texts of economic reports and media outlets in great detail depict the current state of the economy and financial markets, both when it is positive and negative.

Many researchers are now experimenting with textual data, and several have already proved that analysing extensive datasets of texts may enhance our understanding of economic phenomena (Gentzkow et al., 2019). Cerchiello et al. (2017), for instance, determine whether banks are in a distressed state based on news articles. Angelico et al. (2021) have concluded that indexes of inflation expectations that are created using tweets are rather informative. Burri and Kaufmann (2020) have integrated both financial market indicators and news articles to give a comprehensive evaluation of Switzerland's economy during the starting phase of the COVID-19 pandemic. Similar research was conducted in Thorsrud (2020) assessing the state of Norway's economy. Other honourable mentions of text-based economic analysis include Baker et al. (2016), Kalamara et al. (2020), Rambaccussing and Kwiatkowski (2020), and Tetlock (2007).

The specific task of financial crisis identification and prediction has been addressed by only one study using NLP, namely Chen et al. (2023). The approach of this paper served as the basis for my study: the authors have the same objective of identifying a financial crisis and one of their data sources corresponds to mine, mainly the OECD reports. The main method in Chen et al. (2023) is a simple bag-of-words approach, according to which the authors count the number of words in a document and look for those which match with a pre-compiled dictionary of financial terms. This approach, however, tends to lead to the loss of context and the model treats every word in the dictionary the same way. This study aims to resolve the main limitation identified in Chen et al. (2023).

The sophistication of NLP methods has grown exponentially over the last 5 years. Even though simple dictionary-based methods provide adequate, easily interpretable results, researchers are turning more and more to large language models, based on the transformer

architecture (Vaswani et al., 2023), to achieve results of even higher quality. Unlike traditional machine learning algorithms, LLMs offer not only predictions but also the ability to provide explanations. Besides, the pre-trained chat versions of such models allow researchers to get output immediately without the need to train their own transformer model, tailored to a specific task.

LLMs are becoming more prominent in the economics literature, for instance, for simulating the behaviour of humans in hiring and wage negotiation scenarios Horton (2023) or for financial news analysis Chu et al. (2023). Wu et al. (2023) have even created their own BloombergGPT which showed a much better quality on financial tasks than all the existing models prior to that. However, to my knowledge, no study on financial crisis identification or financial crisis prediction has been conducted so far with the use of an LLM.

# Theoretical Framework

The first step to identifying financial crises is to define what constitutes a financial crisis. In the current work, I adhere to the framework suggested in Romer and Romer (2017), where the authors classify episodes of financial distress, reported in OECD Economic Outlook. First, to define the concept of financial distress, the authors themselves turn to Bernanke (1983) and refer to it as the "cost of credit intermediation" which implies the costs of financial institutions to monitor and administer loans and the risks associated with them. When the said cost increases, the supply of credit tends to decrease as the banks face higher expenses due to higher risks. Having defined financial distress, Romer and Romer (2017) affirm that the state of a financial crisis is not binary, they suggest a scale of 0-15 and distinguish between five degrees of crisis severity, basing their descriptions on OECD reports.

1. Credit disruptions (1-3): a slight increase in the cost of credit intermediation which is noteworthy but not likely to lead to any noticeable macroeconomic consequences. Such disruptions may emerge during the period when the country is recovering from a minor crisis.

2. Minor crises (4-6): considerable problems in the financial sector which affect economic output, however, are not recognized as dominant in terms of the effect on the economy's future.

3. Moderate crises (7-9): widespread and serious issues in the financial sector which

affect the country's economic performance overall, but the financial system does not shut down completely. Such a crisis is also characterised by significant government interventions. This definition of a moderate crisis resembles the line between a systemic crisis and a nonsystemic crisis in other works on this topic (Laeven & Valencia, 2013; Reinhart & Rogoff, 2014).

4-5. Major crises (10-12) and extreme crisis (13-15): Romer and Romer (2017) do not specifically give criteria to differentiate between the two extreme degrees, however, they portray them as situations when regular financial intermediation is close to impossible across the whole financial system and government interventions are considered major. OECD analysts tend to use the words "paralysis" and "grave" in their reports to describe such crises.

The task of identifying the degree of severity of a financial crisis is significantly more complex than simply identifying whether a financial crisis is present and only experts in the field with analytical materials and data at their disposal are capable of this task. Hence, the aim of this paper lies only within identifying the existence of a financial crisis with large language models, not its category.

# Empirical Strategy

The textual data - the object of the analysis - is OECD Economic Outlook Reports ("OECD economic outlook: Statistics and projections", n.d.). They have been published semi-annually since 1967, at first only for the seven biggest economies of OECD, and since 1981 for all twenty-four of its members. They all follow a similar outline and thoroughly describe each state's economy across various dimensions, concentrating on major developments.

The initial phase of this research involved the retrieval of reports from the official OECD website through web scraping, followed by the extraction of sections pertaining to individual countries. Given that a significant portion of the older reports were scanned documents, manual copying using Adobe Acrobat's Optical Character Recognition (OCR) was necessary. Due to the presence of numerous figures, tables, and footnotes, the data preparation process was time-consuming. The outcome of this phase was a dataset comprising raw text from 1732 OECD reports spanning the years 1967 to 2012.

The LLM which was chosen as the main tool for the paper was Meta's LLAMA 2-chat

with seventy billion parameters (Touvron et al., 2023). This choice is justified by the fact that at the time, it is considered to be one of the best open-source LLMs, capable of a broad range of tasks (Touvron et al., 2023).

The target variable in the model is binary: its value is one if Romer and Romer (2017) score a crisis from 4 to 15 depending on its severity, and zero in all other cases. The Romer & Romer crisis database (2017) is constrained to the years 1967 to 2012, with no instances of financial distress identified prior to 1990. Thus, the sample which was eventually given to the LLM was limited to years from 1990 to 2012 and consisted of 1104 observations.

To streamline the process of querying the model, the Replicate API was utilised in conjunction with Python code for extracting the model's classification output and explanation ("Replicate", n.d.). Various prompts were experimented with to optimise classification quality metrics, initially on smaller subsets of 85 and 384 reports.

The prompt which showed the best quality:

> *Below is an extract about a certain country from an OECD report. Classify this text as either describing an economy with a financial crisis (category 1) or describing an economy without a financial crisis (category 0).*
>
> *Please focus specifically on the fact that you are looking for a financial crisis, that is, the extract should describe widespread and severe problems in the financial sector which are central to the economy as a whole. You should try to avoid false positives.*
>
> *Your answer should consist of two distinct parts: 1) the category; 2) the explanation for the category.*

The first part of the prompt describes the model's objective.

The second part of the prompt in response to the initial prevalence of false positive answers generated by the model. There were two major reasons for that: 1) only a third of observations in the sample were labelled as financial crisis by Romer and Romer (2017), hence, the sample was unbalanced which skewed the LLM's results; 2) the simple prompts used initially seem to have biassed LLM to look specifically for any indication of a financial crisis, thus, the model mistook any negative events in the economy of a country for a financial crisis. Hence, the definition of a financial crisis from Romer and Romer (2017) was

incorporated into the prompt, accompanied by a cautionary note regarding the propensity for false positive results. The third part of the prompt was supposed to ask the model for the same output format for all answers.

The text above was used as a "system prompt", which is a convenient way of guiding the model's behaviour for all of the messages in a chat ("A guide to prompting Llama 2", 2023). Finally, the "prompt" variable itself consisted only of the text of the report extracted from a sheet compiled earlier.

# Results

As stated in the prompt, for each of the 1104 observations, the model generated a category and an explanation of such categorisation. The current section is split into 1) discussing the binary classification task of identifying the financial crisis, which LLM performed and 2) investigating how the LLM explained its choices.

## Classification Task Results

The sample of reports which was given to the LLM was heavily unbalanced, as it contained only 152 reports describing the period of financial crisis out of 1104 reports total, i.e. approximately only 14%. In addition, 84 of those 152 reports concerned the years between 2007 and 2010, the time of the global financial crisis and its aftermath. The results of the binary classification can be seen in Figure 1.

The model successfully identified a financial crisis in 110 instances and also correctly recognised a lack of financial distress in 902 other reports. Although there are 50 false positive errors, they are not critical, given the objective of the task. When an LLM produces a warning (i.e. a false positive or a true positive answer), the authorities can mitigate the consequences of whatever an LLM noticed by investigating other metrics and either confirming or ruling out the possibility of a crisis. Hence, the most dangerous errors which should be minimized are the false negatives as in this case the central bank is more likely to miss the start of financial distress which will lead to undesirable consequences. Overall, the current model produced 42 negative errors, and their nature will be explored further.

Table 1 demonstrates the different quality metrics which can be calculated from the

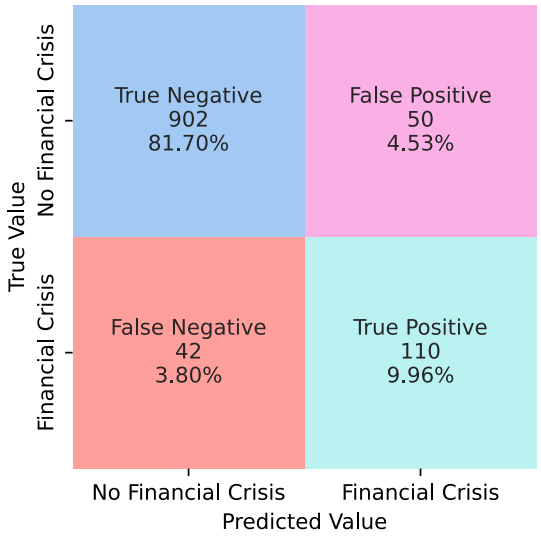


Figure 1: Confusion Matrix for all Observations
Note: own calculations

confusion matrix above. There are two columns: the left one with metrics calculated by usual formulas, those are generally used for any binary classification task; the right column contains the values of the weighted metrics, which depict the quality of the model, accounting for the imbalance of classes. Naturally, these modified metrics hide the performance of the model when it comes to infrequent classes, in our case the observations with financial crises, hence, it is more reasonable to analyse only the left column, although it contains significantly less impressive results (Pedregosa et al., 2011).

The selected quality metrics are widely used for classification tasks and include: 1) *accuracy* - a share of correctly classified instances; 2) *precision* - a share of true positive answers of the model among all positive answers, which allows us to see how well the model is prone to producing false warning signals; 3) *recall* - a share of true positive answers of the model among all actual positive instances, which directly shows the capacity of the algorithm to identify crises and ensure that few relevant instances are missed; 4) *F1-score* - the harmonic mean of precision and recall; 5) *ROC-AUC score* - a measure of the model's ability to distinguish between positive and negative instances, which directly indicates the effectiveness of the algorithm in correctly identifying crises across all threshold levels; the higher ROC-AUC, the better the discrimination capacity of the model.

LLAMA-2 gave a correct answer when identifying a financial crisis based on an OECD economic report in 91.7% of cases. The model successfully recognized that a substantial

Table 1: Quality Metrics

| | **Full Sample** | **Full Sample Weighted Metrics** |
|---|---|---|
| Accuracy | 0.917 | 0.917 |
| Precision | 0.688 | 0.919 |
| Recall | 0.724 | 0.917 |
| F1 | 0.705 | 0.918 |
| ROC-AUC | 0.836 | 0.836 |

Note: own calculations

share of observations do not describe a financial crisis. Nevertheless, some false positives appear, producing a precision score of 0.688. However, as described above, precision should not be the central focus of this classification task, as the consequences from a false negative significantly outweigh the consequences from a false positive. Therefore, special attention should be paid to the recall metric of 0.724, which means that the period of financial crisis was successfully identified by the LLM in 72% of cases. The standard ROC-AUC metric equals 0.836. Both recall and ROC-AUC are comparable enough with the scores obtained in previous research (Bluwstein et al., 2020; Chen et al., 2023; Tölö, 2020), thus, LLAMA-2 even without fine-tuning performs at least as good as classic machine learning models with sophisticated training and validation processes.

## Explanation of LLM's Answers and Error Analysis

A preliminary analysis of why the LLM committed certain mistakes involved looking at the answer matrix for every individual report. Figure 2 demonstrates the results. There are some noticeable patterns in this matrix: while for some countries the model produced one or zero errors (e.g., Australia, Netherlands, Sweden), for other countries there were more than six of them (e.g., Iceland, Italy, United States). Besides, for the latter group Turkey and Luxembourg had mainly false positive errors, while Japan had largely false negative errors. Although the model recognized the Japanese Lost Decade (Hayashi & Prescott, 2000), starting from 1992, it was at times confused by the changing economic landscape. The model was able to detect the significant economic downturn of the 2007-2008, producing a lot of true positive answers, however, in certain cases it seemed to have difficulty distinguishing between minor economic fluctuations and significant economic crisis.

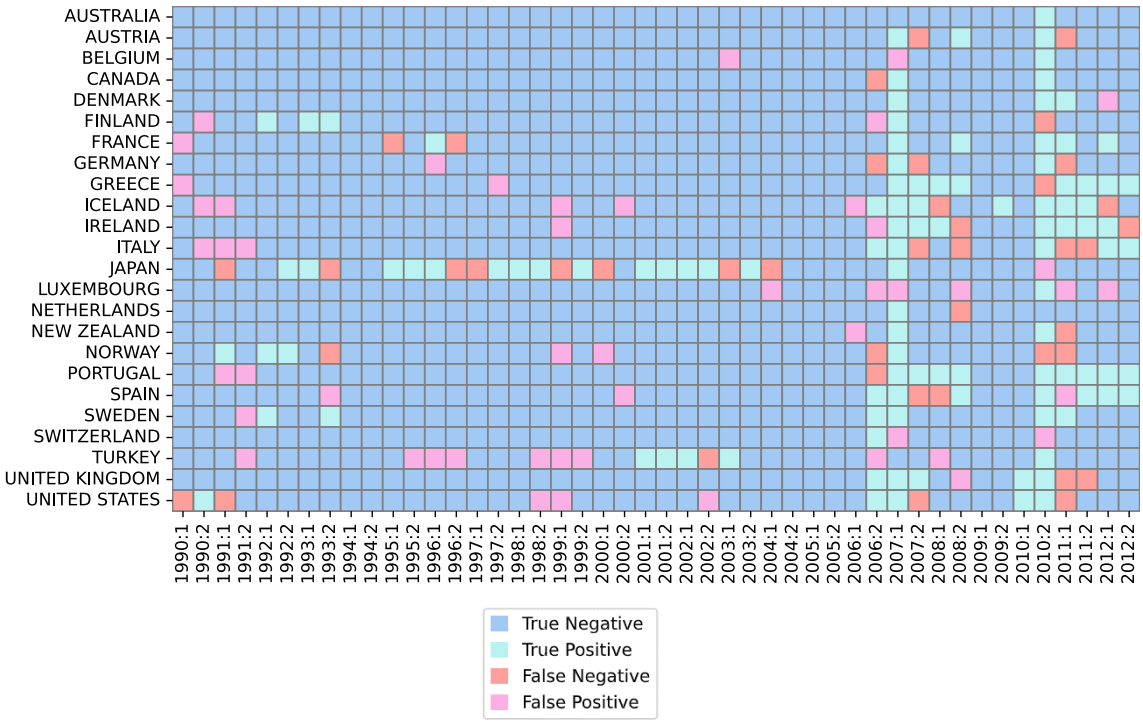


Figure 2: Confusion Matrix for all Observations
Note: own calculations

To demonstrate in detail the model's reasoning behind each type of the answer, the exact text produced by the LLM is provided below along with an extract from the corresponding report.

The analysis of Table 2 provides several crucial insights into how the LLM came to its decisions. For the true negative and positive responses, the model accurately determines the economic context, while paying attention to key indicators such as GDP growth, unemployment, inflation and, most importantly, policies, always specifying if and what fiscal and monetary measures are being implemented. For the false negative answer, the model perceives the signs of recovery and public spending as indicators of economic stability, acknowledging challenges like corporate restructuring within a recovery context. However, this focus on improved activity overshadows the subtle signs of underlying economic fragility. The false positive error arises because the LLM treats high unemployment, the GDP decline and tight policy as markers of a financial crisis. Therefore, the model mistakes severe recessions for financial crises, failing to recognize the qualitative differences between them, and in part the LLM is justified in doing so as it was not fed the methodology which was used in Romer and Romer (2017) to label these crises, hence, the model's opinion on the matter may also be valid.

Table 2: Examples of Model Answers

| Answer | Report Extract | LLM's Explanation |
|---|---|---|
| **True Negative** | Belgium, 2005:1<br><br>Economic growth is projected to slow to 1.3% in 2005 but to rise to 2.4% in 2006 as domestic demand and exports strengthen. Despite a pick-up in employment growth in 2006, the unemployment rate should only ease back to around 8%. Headline inflation is likely to fall to 1.6% as the effects of higher energy prices pass, converging with the underlying rate. Additional consolidation measures will be required to keep the budget in balance. In view of the economic costs of the already high tax burden, these should focus on expenditure restraint. | The extract describes an economy that is experiencing a slowdown in growth, but there is no mention of a financial crisis. The economy is projected to grow at a rate of 1.3% in 2005 and 2.4% in 2006, with private consumption and business investment expected to strengthen. While the unemployment rate is expected to remain relatively high at around 8%, there is no indication of widespread job losses or financial instability. Inflation is also expected to fall to 1.6% in 2006, which suggests that there is no significant inflationary pressure. Additionally, the government's budget position is expected to be in balance, with additional consolidation measures planned to keep it that way. Overall, the extract does not describe an economy facing a financial crisis. |
| **True Positive** | Iceland, 2006:2<br><br>There are increasing signs that a change in foreign investor sentiment early this year and further policy tightening has set in motion an overdue adjustment process. The economy is projected to contract in the next few quarters. Nonetheless, economic imbalances will remain substantial for some time. Thus, the challenge for policymakers will be to ensure that steady progress is made in unwinding them lest that result in a hard landing for the economy. Continued high household and business sector indebtedness and market nervousness make a significant correction likely. Further tightening of already tight monetary and fiscal policies is likely to be needed in combination with further inflation to target and enhance credibility. | The text mentions a change in foreign investor sentiment, a contraction in the economy, and a large current account deficit which are clear indicators of a financial crisis. Additionally, the text notes that policy tightening has been implemented to address economic imbalances, but these efforts might not lead to a complete resolution and could still result in a hard landing for the economy. The mention of high household and business sector indebtedness and market nervousness indicates significant financial instability. Further tightening of monetary and fiscal policies is likely as part of the measures needed to handle the financial crisis, which could exacerbate the situation. |
| **False Negative** | Japan, 2000:1<br><br>Clear signs of improved economic activity in recent months have eased earlier concerns about a deflationary spiral, despite statistical ambiguities pertaining to the second half of 1999. The economy now appears to be on the path of a cyclical recovery, boosted in the first half of this year by public works spending and by policy-induced housing construction which will continue till next spring. Some temporary strengthening of private consumption is likely, but its underlying trend may remain weak as continuing corporate restructuring weighs on growth of wage incomes. Output growth is projected to reach an annual rate of 3 per cent in 2000. Uncertainties remain as to how solidly based is the on-going recovery. | The extract describes an economy that is experiencing a cyclical recovery, with improved economic activity, rising profits, and a pick-up in investment. While there are some remaining challenges, such as corporate restructuring and over-indebtedness, the overall picture is one of a recovering economy. There is no mention of a financial crisis or any widespread and severe problems in the financial sector that are central to the economy. Instead, the focus is on the government's efforts to support the economy through monetary and fiscal policy, and the potential for future growth. Therefore, this extract does not describe an economy with a financial crisis and should be classified as category 0. |
| **False Positive** | Spain, 1993:1<br><br>The Spanish economy went into recession in mid-1992. This was accompanied by a marked rise in the unemployment rate, to 21.7 per cent in early 1993. GDP declined further in the first quarter of 1993, contributing to a fall in inflation and the current external deficit. The policy stance remains tight, but because of large increases in the cyclical component of the budget deficit and in interest payments, the general government deficit may increase to 5.4 per cent of GDP in 1993. Supported by faster net export growth following the 1992 and 1993 devaluations of the peseta, activity is projected to pick up slowly in the second half of 1993. | The text describes a severe economic downturn in Spain, with high unemployment, a large budget deficit, and a depreciating currency. The economy is experiencing a recession, with GDP declining and inflation rising. The government's policy stance is tight, with measures to reduce the budget deficit and combat inflation, but the economy is still struggling. The text also mentions that the peseta was frequently under speculative pressure, leading to a devaluation, and that the current external deficit is high. All of these factors indicate that Spain is experiencing a financial crisis. Therefore, the category is 1 (Financial crisis). |

While Table 2 allows us to look at individual examples, Table 3 provides a broad overview of the model's errors by looking at key words which are characteristic of the model's explanations in the corresponding category. To determine the key words, the following methodology was adopted. All the explanations which were given by the LLM were split into four groups in accordance with the LLM's answer, and for each group a merged text was compiled. The four texts were then tokenized, in this case simply split by words, and the tf-idf score was calculated for each word. The tf-idf score helps to identify important, frequently occuring words, which are unique to the specific text in a corpus (Qaiser & Ali, 2018). Due to the formula for tf-idf, the term frequency may significantly alter its value if a particular word is encountered unproportionally more often. Thus, Table 3 shows the words ranked by their tf-idf score, however, some words were removed because of the lack of the connection between their meaning and the economic context of the task. Those words included "does", "mention", "including", "text", "factors", "additionally", among others.

Table 3: Interpretation of the LLM's Explanations by Type of Error

| | **True Negative** | **False Negative** | **True Positive** | **False Positive** |
|---|---|---|---|---|
| **1** | growth | recovering | facing | significant |
| **2** | widespread | recession | significant | facing |
| **3** | problems | recovery | banks | account |
| **4** | inflation | positive | conditions | deficit |
| **5** | strong | risks | contraction | inflation |
| **6** | positive | potential | decline | high |
| **7** | stable | banking | fiscal | challenges |
| **8** | robust | policies | credit | deterioration |
| **9** | severe | consolidation | unemployment | wage |
| **10** | low | plans | domestic | borrowing |
| **Average Compound Sentiment Score for the report** | | | | |
| | 0.86 | 0.77 | 0.23 | 0.61 |
| **Average Compound Sentiment Score for the explanation** | | | | |
| | -0.39 | -0.38 | -0.85 | -0.80 |

*Note: own calculations*

The analysis of these words provides even more insights how the LLM rationalized its

choice of a particular label for each of the reports. For true negative responses, we observe "growth", "strong", "positive", "stable" and "robust" as clear indicators of a healthy economy without any financial distress. All the words which could be described as having a negative sentiment ("problems" or "inflation") were actually used as part of set phrases, such as "no widespread problems", "no severe problems", "low inflation", thus, the model was underlining the lack of a financial crisis.

The occurrence of false negative results is self-explanatory, looking at Table 3. Those errors were mainly associated with the reports where the LLM acknowledged the economies in the process of "recovery". The country is coming out of a "recession", and although "potential risks" exist, the outlook on the "banking" industry is "positive" with future "policies" and "plans" set in place. This is the portrait of an explanation which the LLM created for these reports, and it is extremely similar to the case of Japan in Table 2. The model focuses too much on the positive and overlooks the fact that the crisis has not yet come to an end.

For the true positive answers, the model describes an economy "facing significant challenges" with much "contraction" and "decline" and a special focus being placed on "banks" and "unemployment". Naturally, the model argues its positive answer very similarly for the false and true answers. It finds "significant challenges", "current account deficit" and "high inflation" as the main markers of an ongoing financial crisis for these two groups of answers.

In addition to the analysis of the td-idf scores in the LLM's explanations, a validation check was conducted. A compound sentiment score was calculated for all the reports in the sample and then a mean score was determined for each of the four groups of answers. This score was calculated using the nltk Python library (Bird et al., 2009) and is a single measure which shows whether the text is negative (value of -1), or the text is positive (value of 1). Naturally, the true negative reports are the most positive (0.86) and the false negative reports are not far behind (0.77). The reports which the model correctly identified as crises are significantly more negative in sentiment than all other reports, whereas when the model identified a non-existent crisis, the reports were closer in sentiment to those without any distress.

Moreover, the LLM's explanations themselves were analysed in terms of sentiment. Although all of them were much more negative than the reports, there is a clear difference between the LLM's negative answer (scores of -0.38 and -0.39) and its positive answer (scores of -0.8 and -0.85). All of the analysis above provides further clues into the LLM's decision-making process and its mistakes.

# Conclusion

## Discussion

The results of this study demonstrate the potential of Large Language Models (LLMs) in identifying financial crises from OECD economic reports. The performance of the model turned out to be on par with the traditional machine learning models applied to textual data, suggesting that LLMs can serve as additional tools in economic forecasting and crisis detection without requiring extensive fine-tuning.

The analysis of the LLM's explanations provided insights into its decision-making process. Both the investigation of individual cases as well as the broad overview analysis lead to similar conclusions about the logic which the LLM followed to make its decisions. It excels in identifying key economic indicators, such as GDP growth, unemployment, inflation, and policy measures, and linking them to broader economic contexts, however, it overemphasizes positive indicators in the presence of recovery signals, or describes severe economic distress, but does not suggest systemic financial instability typical of a crisis.

The country-specific analysis revealed that the model performs well for certain countries, such as Australia, Netherlands, and Sweden, but struggles with others, like Iceland, Italy, Japan, Turkey and the United States. The error patterns suggest that the model may be influenced by country-specific economic contexts and policy responses. In the context of financial crises, the study's findings have significant implications for central banks and policymakers. The ability to accurately identify financial crises is crucial for timely and effective policy responses. The use of LLMs can enhance the speed and accuracy of crisis detection, allowing policymakers to respond more quickly to emerging crises. However, the study's results also highlight the need for caution in relying solely on LLMs for crisis detection. The model's tendency to overemphasize positive indicators and mistake severe economic distress for financial crises underscores the importance of human judgment and expertise for robust and reliable economic analysis and policy-making.

Furthermore, the study raises important ethical considerations for central banks and other regulatory bodies as the controlled use of LLMs is crucial in ensuring fairness and accuracy in economic policy-making. It is essential to address the potential biases that may arise with the use of LLMs as they can result in the disproportionate identification of crises in particular countries or sectors. Such biases may emerge due to overrepresentation or underrepresentation of certain regions or economic activities in the training data. If the

LLM is biased towards identifying crises in certain countries or sectors, its use may lead to uneven policy responses and amplify existing economic disparities.

To mitigate such a risk, policymakers should implement robust measures to ensure the representativeness of the training data as well as comprehensive regulatory frameworks which involve transparent methodologies and continuous monitoring of the LLM's performance. Besides, policymakers should engage in multi-stakeholder consultations, including economists, data scientists and philosophers to develop and refine the guidelines on the use of LLMs for complicated economic analysis.

## Limitations

Since the current research is one of the first studies on the use of LLMs to solve a sophisticated abstract task from the domain of economics, it is essential to discuss its limitations to provide a comprehensive understanding of the findings.

Firstly, the most substantial concern involves the fact that the main source of the textual data which was fed to the LLM was the OECD Economic Outlook Reports. These reports are published semi-annually with a significant time gap which does not at all eliminate the initial concern of untimely crisis identification. The better alternative would be to feed the LLM thousands of news pieces which are produced daily in order to get a constantly updating indicator. However, this does not diminish the value of the current study due to the fact that it is still the first one to look at the possibility of using LLMs for crisis identification.

Secondly, the sample consisting of OECD reports was chosen due to the presence of labels which were carefully done in Romer and Romer (2017) for each of the observations. This significantly limited the number of countries and crisis periods under investigation and led to a heavily unbalanced sample.

Finally, there are numerous to modify the parameters of the LLM to achieve the highest accuracy, including fine-tuning and few-shot learning. Fine-tuning was initially attempted by the author to create a model which would be specifically tailored to the task of crisis identification, however, due to computational constraints, this was not effective. Few-shot learning, i.e. showing several examples of how to categorize reports to the LLM, is also bound to increase the quality metrics, but it was not implemented due to the large size of the reports and a limited number of tokens which could be fed to the model at a time.

## Further Research

Further research can be built upon the limitations of this study. More specifically, a different dataset should be explored, ideally in a different domain, such as news articles. Few-shot learning and fine-tuning may also enhance the model's ability to differentiate between severe recessions and financial crises. Besides, methodologies should be developed to integrate LLMs with traditional numerical economic models and expert judgment. Also, further investigation into the LLM's interpretability and explanation mechanisms is required. Understanding how the model arrives at its decisions and improving the transparency of its reasoning process is essential for economic policy-making. Finally, it is crucial to consider the ethical implications of relying on LLMs in crisis detection and to ensure that their use is safe, fair and accountable.

# References


A guide to prompting Llama 2. (2023, August). https://replicate.com/blog/how-to-prompt-llama#system-prompts

Alessi, L., & Detken, C. (2018). Identifying excessive credit growth and leverage. *J. Fin. Stab.*, *35*, 215–225.

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2021). Can we measure inflation expectations using twitter? *SSRN Electron. J.*

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Q. J. Econ.*, *131*(4), 1593–1636.

Bernanke, B. (1983). *Non-monetary effects of the financial crisis in the propagation of the great depression* (tech. rep.). National Bureau of Economic Research. Cambridge, MA, National Bureau of Economic Research.

Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? *J. Fin. Stab.*, *45*(100693), 100693.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*.

Bluwstein, K., Buckmann, M., Joseph, A., Kang, M., Kapadia, S., Kapadia, S., & Simsek, Ö. (2020). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. https://doi.org/10.2139/SSRN.3520659

Brave, S., & Butters, A. R. (2012). Diagnosing the financial system: Financial conditions and financial stress. *& International Journal of Central Banking*.

Burri, M., & Kaufmann, D. (2020). A daily fever curve for the swiss economy. *Schweiz. Z. Volkswirtsch. Stat.*, *156*(1), 6.

Cerchiello, P., Giudici, P., & Nicola, G. (2017). Twitter data models for bank risk contagion. *Neurocomputing*, *264*, 50–56.

Chen, M., DeHaven, M., Kitschelt, I., Lee, S. J., & Sicilian, M. J. (2023). Identifying financial crises using machine learning on textual data. *Int. Fin. Discuss. Pap.*, (1374), 1–40.

Chu, Z., Guo, H., Zhou, X., Wang, Y., Yu, F., Chen, H., Xu, W., Lu, X., Cui, Q., Li, L., Zhou, J., & Li, S. (2023). Data-centric financial large language models. *ArXiv*, *abs/2310.17784*. https://doi.org/10.48550/arXiv.2310.17784

Drehmann, M., & Juselius, M. (2014). Evaluating early warning indicators of banking crises: Satisfying policy requirements. *Int. J. Forecast.*, *30*(3), 759–780.

Duca, M. L., & Peltonen, T. A. (2013). Assessing systemic risks and predicting systemic events. *J. Bank. Financ.*, *37*(7), 2183–2195.

Eichengreen, B. J., & Rose, A. K. (1998). Staying afloat when the wind shifts: External factors and emerging market banking crises. *SSRN Electron. J.*

Fouliard, J., Howell, M. J., & Rey, H. (2020). Answering the queen: Machine learning and financial crises. *SSRN Electronic Journal*. https://doi.org/10.3386/w28302

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *J. Econ. Lit.*, *57*(3), 535–574.

Hanson, S. G., Kashyap, A. K., & Stein, J. C. (2010). A macroprudential approach to financial regulation. *SSRN Electron. J.*

Hayashi, F., & Prescott, E. (2000). The 1990s in Japan: A Lost Decade. *Review of Economic Dynamics*, *5*, 206–235. https://doi.org/10.1006/REDY.2001.0149

Holopainen, M., & Sarlin, P. (2015). Crisis modeler: A tool for exploring crisis predictions. *2015 IEEE Symposium Series on Computational Intelligence*.

Horton, J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4413859

Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M., & Taylor, A. (2017). *The Rate of Return on Everything, 1870–2015* (tech. rep.). https://doi.org/10.3386/w24112

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making text count: Economic forecasting using newspaper text. *SSRN Electron. J.*

Kaminsky, G., Lizondo, S., & Reinhart, C. M. (1998). Leading indicators of currency crises. *Staff papers - International Monetary Fund*, *45*(1), 1. https://doi.org/10.2307/3867328

Knedlik, T., & Von Schweinitz, G. (2012). Macroeconomic imbalances as indicators for debt crises in europe. *J. Common Mark. Stud.*, *50*(5), 726–745.

Laeven, L., & Valencia, F. (2013). Systemic banking crises database. *IMF Econ. Rev.*, *61*(2), 225–270.

Laeven, L., & Valencia, F. (2020). Systemic banking crises database II. *IMF Econ. Rev.*, *68*(2), 307–361.

Lee, S. J., Posenau, K. E., & Stebunovs, V. (2020). The anatomy of financial vulnerabilities and banking crises. *J. Bank. Financ.*, *112*(105334), 105334.

OECD economic outlook: Statistics and projections. (n.d.). https://www.oecd-ilibrary.org/economics/data/oecd-economic-outlook-statistics-and-projections_eo-data-en

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qaiser, S., & Ali, R. (2018). Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*. https://doi.org/10.5120/IJCA2018917395

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *Int. J. Forecast.*

Reinhart, C. M., & Rogoff, K. S. (2009). The aftermath of financial crises. *Am. Econ. Rev.*, *99*(2), 466–472.

Reinhart, C. M., & Rogoff, K. S. (2014). Recovery from financial crises: Evidence from 100 episodes. *Am. Econ. Rev.*, *104*(5), 50–55.

Replicate. (n.d.). https://replicate.com/

Romer, C. D., & Romer, D. H. (2017). New evidence on the aftermath of financial crises in advanced countries. *Am. Econ. Rev.*, *107*(10), 3072–3118.

Sachs, J. D., Tornell, A., Velasco, A., Calvo, G. A., & Cooper, R. N. (1996). Financial crises in emerging markets: The lessons from 1995. *Brookings Pap. Econ. Act.*, *1996*(1), 147.

Samitas, A., Kampouris, E., & Kenourgios, D. (2020). Machine learning as an early warning system to predict financial crisis. *International Review of Financial Analysis*, *71*(100). https://doi.org/10.1016/j.irfa.2020.10150

Sarkar, A., & Patel, S. A. (1998). Stock market crises in developed and emerging markets. *SSRN Electron. J.*

Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2014). Developing an early warning system to predict currency crises. *Eur. J. Oper. Res.*, *237*(3), 1095–1104.

Shin, H. S. (2011). Bank for international settlements. monetary and economic department, & chosŏn ŭnhaeng. *Macroprudential regulation and policy*, *60*, 5–15.

Tanaka, K., Kinkyo, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Econ. Lett.*, *148*, 118–121.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *J. Finance*, *62*(3), 1139–1168.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *J. Bus. Econ. Stat.*, *38*(2), 393–409.

Tölö, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, *49*(100). https://doi.org/10.1016/j.jfs.2020.100746

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, *abs/2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *ArXiv*, *abs/2303.17564*. https://doi.org/10.48550/arXiv.2303.17564