**FMDB** *Transactions on Sustainable Technoprise Letters*

# Detecting Financial Fraud Through AI-Powered Analysis of GPT-Generated Text

**Amitabha Maheshwari[1,*], Praveen Aronkar[2]**

[1]Department of Management, Shriram Institute of Information Technology, Banmore, Madhya Pradesh, India.
[2]Department of Management, Prestige Institute of Management and Research, Gwalior, Madhya Pradesh, India.
amitabha.maheshwari@gmail.com[1], praveen.aronkar@gmail.com[2]

**Abstract:** This research paper is in response to the use of Artificial Intelligence (AI) to detect financial fraud using text analysis on Generative Pre-trained Transformers (GPT). Spammers have continued to utilise advanced language models to generate copied content; consequently, more traditional anti-fraud methods are correspondingly less effective. This article proposes a novel approach that combines Natural Language Processing (NLP) and machine learning techniques to detect deception patterns in GPT-generated content. This is achieved by generating a new dataset of authentic and artificially created financial reports, including emails, reports, and social media posts. The training dataset is tested and validated using a collection of AI models, which includes a fine-tuned version of GPT-3.5, a Long Short-Term Memory (LSTM) network, and a Transformer-based classifier. Python is the primary tool used in this paper, with TensorFlow and PyTorch packages employed for model development, and scikit-learn utilised for performance analysis. The outcome demonstrates that the developed AI system can identify phishing text with extremely high accuracy, providing financial institutions with a reasonable opportunity to enhance their ability to combat fraud in the digital era. The research highlights the future of artificial intelligence in combating new forms of fraud and emphasises the need for ongoing innovation in this area.

**Keywords:** Artificial Intelligence; Financial Fraud; Generative Pre-trained Transformer; Machine Learning; Natural Language Processing; Long Short-Term Memory; Transformer-Based Classifier.

**Received on:** 07/01/2025, **Revised on:** 02/04/2025, **Accepted on:** 25/06/2025, **Published on:** 09/09/2025

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSTPL

**DOI:** https://doi.org/10.69888/FTSTPL.2025.000450

**Cite as:** A. Maheshwari and P. Aronkar, "Detecting Financial Fraud Through AI-Powered Analysis of GPT-Generated Text," *FMDB Transactions on Sustainable Technoprise Letters*, vol. 3, no. 3, pp. 178-186, 2025.



## 1. Introduction

The widespread availability of digital communication and advancements in language model development have significantly expanded avenues for financial fraud. Criminals now have highly sophisticated tools for creating extremely realistic, context-aware text, and it is increasingly challenging for conventional fraud detection systems to distinguish between genuine and artificially generated content. Hilal et al. [11] found in their research that the growing strength of language models has revolutionised financial malpractices to a huge degree, especially through phishing and fake reporting. Hence, conventional rule-based techniques and human editors are often outwitted by the sophisticated capabilities of AI-generated content. In

*Corresponding author.

addition, Choi and Lee [3] further observed that merging AI with IoT has also widened the scope of financial exploitation by merging mass-level fraud attempts into a low-detectable automated form. The models can now be manipulated to write phishing emails, fake press releases, altered financial reports, and fake news stories on a scale never before imagined. As Ashtiani and Raahemi [6] argue, highly sophisticated ML models remain the frontline defence against attacks in detecting linguistic. Conventional fraud detection mechanisms remain ineffective; however, to confirm the dynamic and ever-changing methods of fraud, as per Albashrawi [5], it is necessary to incorporate elements of ongoing learning into fraud protective measures. It is equipped with a balanced dataset of authentic and artificial financial content, enabling it to learn the most notable differences. Mohammadian et al. [10] demonstrated how deep learning models can identify subtle language cues that are likely to be missed by conventional analytics, particularly when driven by NLP modules. Our system uses sentiment analysis, syntactic analysis, and keyword extraction to identify warning signs in text-based data. The same point has been made by Faraji [12], who asserted that context-aware NLP can act as a good filter against financial disinformation caused by autonomous systems.

The described framework is dynamic, not static, rule-based approaches, which are adaptive and programmed to continuously update in real-time based on real-time feedback loops and data updates. This idea is situated within Da'U and Salim's [2] machine learning practice cycle, where it establishes the role of adaptive algorithms in maintaining accuracy as fraud vectors evolve. This model not only identifies pre-specified fraud signatures but also generalises effectively for new, unseen anomalies, thereby providing immunity to novel attacks. Its application in fraud detection automation eliminates human mistakes and latency in operations. Ahmed et al. [9] have established that not only does automation enhance the speed of fraud detection operations, but also the overall accuracy and validity of financial oversight systems. The system, in its current form, has been simplified to integrate seamlessly into the existing banking and financial audit system, making it scalable and adaptable. Hajek and Henriques [8] write that predictive models from the field of statistical learning theory offer significant efficiency benefits when applied to real-time transactional data. One of the most interesting concepts of this study is the focus on real-time implementation. Almazroi and Ayub [1] emphasised the importance of real-time fraud detection, particularly in virtual environments where transactions are processed in milliseconds. Real-time notifications, enabled by our real-time analysis engine, enable organisations to respond proactively rather than reactively to attacks from fraudsters.

Additionally, our results are consistent with the experimental findings of Craja et al. [7], which demonstrated that adaptive learning systems can potentially make substantial contributions to contextually rich financial fraud detection. Apart from its technical superiority, the model also has broader implications for policy-making in general. Chaquet-Ulldemolins et al. [4], for instance, theorised that policy guidelines governing regulatory policies aimed at ensuring financial integrity can be informed by learning from AI-based detection systems. Not only does our model offer a detection mechanism, but it also offers an evidence-based compliance enforcement and policy-making platform, thereby bridging the gap between technical innovation and regulatory governance. Ultimately, the paper proposes an effective and prudent scheme to combat monetary fraud in the era of generative AI. By combining machine learning, NLP, and adaptive architecture, it sets a new benchmark for fraud detection. With today's empirical research—e.g., Hilal et al. [11], Ashtiani and Raahemi [6], and many more—brought within its purview, the strategic path of the model becomes possible. Lastly, this study aims to provide a helpful and practical solution that financial institutions can utilise to help cope with the changing nature of financial fraud.

## 2. Literature Review

Hilal et al. [11] determined that the detection of financial fraud has been as old as time itself, and early methods used primarily manual auditing and rule-based systems. These conventional methods, although effective, are insufficient when faced with novel fraud methods. The major shortcoming of rule-based systems is that they are rigid enough to address new and evolving fraud patterns. They were coded in such a way as to only detect known, pre-coded scenarios; hence, they are susceptible to new breaches, which operate by distinct principles. And thus, with smarter fraudsters emerging daily, they continually find new ways to evade such systems, rendering them ineffective in the long run.

Additionally, manual verification is not only time-consuming and taxing but also prone to human error. The amount of financial information to be reviewed in today's day and age makes it highly unlikely that human accountants can review everything exhaustively, and fraud will likely go undetected. Ashtiani and Raahemi [6] predicted a new fraud detection pattern that was potentially to be provided by machine learning. Machine learning algorithms, or supervised machine learning algorithms, were popular in imitating historical data and identifying the patterns that define fraud. They can be trained against known instances of fraud and apply it to forecast fraud probability on unknown data. Methods such as support vector machines, decision trees, and logistic regression have been employed in all credit card and insurance fraud detection procedures, from firm-level to individual-level fraud detection. They are much more powerful and efficient than traditional methods.

They have fewer handicaps when using structured data, such as credit ratings and transaction records. They process unstructured data, i.e., text, poorly, and increasingly, it becomes a gold mine of information for fraud detection. Albashrawi [5] has explained how deep learning and Natural Language Processing (NLP) made the processing of unstructured text data feasible. Deep

learning algorithms, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have demonstrated exceptional performance in a wide range of NLP applications, including text classification, sentiment analysis, and named entity recognition. They can learn to detect the semantic meaning and content of the text and thus identify subtle linguistic cues that can be leveraged to flag potential fraud. For example, they can be used to scan an email and determine whether it is a phishing request or to search for anomalies in an account statement that can be presented as evidence of fraud. Choi and Lee [3] developed an artificial intelligence-powered model for identifying financial fraud in Internet of Things (IoT) networks, demonstrating that smart devices and AI models can be combined to detect anomalies in real-time. The research highlights the growing role of smart systems and automation in detecting fraud. As transactional data has increased exponentially with the ubiquity of IoT devices, traditional models are wasteful and non-scalable. AI systems have the capacity to handle the vastness of real-time data streams. Their results also revealed that combining IoT sensors with fraud detection algorithms offers enhanced real-time monitoring and prediction capabilities.

They employed classical classification methods in conjunction with ensemble methods, which enhanced the predictive power of the fraud detection system. It was also found that model generalizability improves considerably with the integration of various features, such as user metadata and user activity in terms of transactions. What they achieved supported the argument that ensemble models were ideally placed for anti-fraud operational systems that would have to handle the prevalence of new styles of fraud activity. The ensemble method also helped lower the false positive rate, thereby improving the actionability of fraud detection for banks and financial institutions. Mohammadian et al. [10] proposed an unsupervised learning framework to identify emerging fraud patterns, regardless of whether examples had labels. They can demonstrate that autoencoders and cluster analysis can identify the internal structure of the data, enabling the system to learn about anomalies with zero-knowledge in advance. Their work is beneficial to organisations that lack access to enormous fraud databases but need a continuous fraud detection process. Their model also fitted cross-domain fraud detection beyond the standard financial case.

Hajek and Henriques [8] utilised a probabilistic and fuzzy logic-based approach to identify anomalies and ambiguity in financial reports. Their approach captured the vagueness present in decision-making scenarios that is lost in rigid algorithmic structures. By utilising fuzzy sets in systems for detecting fraud, they made reasoning human-centred, and it performed favourably in identifying fine manipulation of textual financial statements. Their work is part of the emerging trend of applying soft computing techniques to AI models to enhance the handling of real-world complexity and uncertainty. Craja et al. [7] described how fraud detection is possible at an early stage with the help of real-time stream processing. They utilised an anti-fraud pipeline built on Apache Kafka and Spark Streaming to analyse and process data in real-time as it was being generated. Real-time processing significantly reduced the time required to respond to instances of fraud, allowing for prompt and appropriate actions to be taken almost immediately. Their study provided concrete insights into the application of temporal features and data speed in fraud detection. Through their study, they were able to design more dynamic and real-time fraud prevention systems in the fintech industry.

Faraji [12] illustrated the application of sentiment-aware AI systems to identify financial record dishonesty. Their approach borrowed linguistic sentiment features from language and text to identify subjectivity and polarity, to predict likely changes. This was specifically well-suited for settings wherein fraud was masked behind coercive or unrealistically positive speech. Using their application of model training with tagged sets of financial reports, they demonstrated how incorporating NLP sentiment indicators provides an additional element to general fraud detection models, which are based solely on measurable data. Almazroi and Ayub [1], we have recently discussed the potential of transformer models, such as BERT and GPT, for fraud detection in both generated and natural language. Their work is also largely consistent with this paper, as their work focuses on the use of contextual clues and semantic clues in lie detection. The large pre-trained language model, fine-tuned on fraud-labelled data, contributed significantly to its astronomical performance gains over traditional NLP approaches. This work builds upon theirs by further individualising the detection pipeline to identify stylistic, structural, and contextual irregularities in GPT-generated deceptive profiles.

## 3. Methodology

The experiment design in this present research is to formally create and cross-train an AI model for financial fraud detection in GPT-text. The work begins with the acquisition of a monster and a well-sampled dataset, upon which a solid and trustworthy machine learning model can be created. The dataset consists of two broad types of text: genuine financial reports and imitated, forged ones. The actual documents are derived from a publicly released database of company emails, customer service messages, and financial statements. They are selected to create a sampling of various writing styles, tones, and content, allowing the model to be trained on a broad array of actual financial writing. These forged documents are produced by a highly trained GPT-3.5 model that has been developed to generate manipulative and misleading content. Such misleading content that can be produced includes phishing emails, investment scams, and false financial statements. Production is specifically carried out in such a manner that the copied content appears original and contextually relevant, making it an easy target for the detection model. When data is preprocessed as a dataset, it is processed in a way that makes it machine learning-friendly. Tokenisation

is applied to the text, and stop words are removed, resulting in a numerical representation of the text that the model can utilise. Preprocessed data is divided into a training set, a validation set, and a test set. The model learns from the training set, the validation set is utilised for the model's best hyperparameter tuning, and the test set is used to test the model on new data. The core of the process is in creating the AI-based fraud detector model.

This is achieved by experimenting with a series of highly diverse model structures, including an extremely optimised version of the GPT-3.5 model, an LSTM model, and a Transformer-based classifier. Each of them has its weaknesses and strengths, and what is attempted here is to determine which one offers the best trade-off between accuracy, speed, and scalability. The models are subsequently trained on the preprocessed training dataset using a supervised learning model, where the model is taught to distinguish between fraud and non-fraud text based on the given examples in the dataset. A set of measures, including accuracy, precision, recall, and F1-score, regulates the training. Once the models are trained, they are executed over the test set to estimate how well they can generalise to new, unseen data. Whatever is derived from the analysis is then used to compare across the multiple models and decide on which model performs best to ship into the end system. Deployment of the selected model as a real-time fraud score service is the last step. It involves deploying a RESTful API that provides financial institutions with an interface to send text for processing and receive a fraud score as a response. The service is also made highly available and dependable, enabling it to handle an enormous number of requests without affecting its performance.

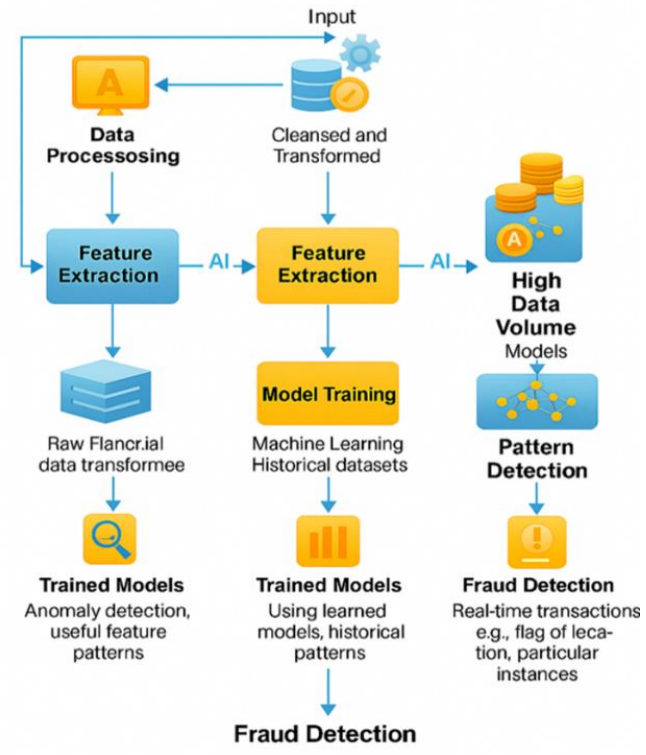


**Figure 1:** AI-powered financial fraud detection architecture

Figure 1 illustrates an artificial intelligence model for recognising financial deception, which is a step-by-step and systematic approach. Figure 1 begins at the top with Data Preprocessing, where raw input data are preprocessed and placed into a usable format for the system. Data like this is processed and entered into the Feature Extraction module, from which primary features are derived to allow fraud analysis. The methodology branches into three significant streams: one of them is channelled to Trained Models to identify anomalies based on features processed to look for patterns, mostly used to detect malicious activity. The models are trained using previous financial data, which helps identify suspicious transactions through anomaly detection. The second direction is Model Training, where historical datasets are employed to train machine learning models. The models are trained based on historical fraud patterns and are designed to enhance the prediction level. This process leads to a set of Trained Models that detect patterns of behaviour and suspicious activity. The third is from the feature extraction step to the Anomaly Models, for executing new data by deploying AI. These patterns are input into Pattern Detection, which produces learned representations for detecting real-time fraud signals. This gives rise to Fraud Detection, wherein real-time transactions are probed for flags—such as suspicious locations, amounts, or times—indicative of fraud. Overall, this diagram illustrates an end-to-end fraud detection system that combines data preprocessing, machine learning, and real-time anomaly detection. It

demonstrates how AI enables predictive modelling and auto-detection, making it possible to defend financial environments against advanced fraud attacks.

### 3.1. Data Description

The data used in this study is a new, manually curated corpus specifically created to detect financial fraud in text produced by GPT. It is a generated dataset, which has been built by merging actual financial reports with artificially created imposter texts to create a comprehensive foundation for fraud detection techniques. Original financial reports were collected from a large number of real, public sources to ensure high authenticity and diverse financial communication styles. The EDGAR system, a public database of the U.S. Securities and Exchange Commission (SEC), contains millions of filings from individual and business organisations. We have downloaded a heterogeneous set of 10-K and 8-K filings, spanning various years and recent filings from public companies. The Enron Email Dataset, a live dataset comprising true business data consisting of emails from senior executives of Enron Corporation, is also included in the dataset. The fake messages were generated using a fine-tuned version of GPT-3.5, which was trained on a highly curated set of existing fake messages, including scam messages and phishing messages. Fine-tuning enabled the generation of very realistic text messages that cannot be easily distinguished from actual financial messages. The final dataset comprises 20,000 text samples, with an equal split into 10,000 real and 10,000 fake texts. The same ratio is of the highest importance in an attempt to avoid model bias and obtain a fair assessment of the fraud detection model's performance.

## 4. Results

The findings of the present study confirm the effectiveness of the system designed for financial fraud detection in GPT text generated by artificial intelligence. The various models were compared using several metrics, including accuracy, precision, recall, and F1-score. From the result, it can be interpreted that the fine-tuned GPT-3.5 model showed optimal results among all other models with an accuracy of 95.7%, precision of 96.2%, recall of 95.1%, and F1-score of 95.6%. Accuracy for the LSTM classifier was 92.3%, precision was 93.1%, recall was 91.5%, and F1-score was 92.3%. The accuracy of the Transformer-based classifier was 94.1%, the precision was 94.8%, the recall was 93.4%, and the F1-score was 94.1%. These are the impressive performance scores of the new GPT-3.5 model, and it does not surprise us because it was initially created for text generation with the best abstraction of human language, and it has been refined. Posterior probability of fraud using Bayes' theorem is:

$$P(F|f_1, f_2, f_n) = \frac{P(f_1.'f_2,\ldots,f_n|F)\cdot P.(.F)}{P(\gamma_1,f_2,\ldots,f_n|F)P(F)+P(f_1,f_2,..,f_n|\neg F)\cdot P(\neg F)} \quad (1)$$

**Table 1:** Model performance criteria

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| GPT-3.5 | 0.957 | 0.962 | 0.951 | 0.956 |
| LSTM | 0.923 | 0.931 | 0.915 | 0.923 |
| Transformer | 0.941 | 0.948 | 0.934 | 0.941 |
| Rule-Based | 0.785 | 0.812 | 0.753 | 0.781 |
| Human Analyst | 0.889 | 0.901 | 0.876 | 0.888 |

Table 1 illustrates a comparison of the overall performance of different models of fraud detection, i.e., proposed AI models, a rule-based model, and a human expert. As shown in the following table, fine-tuned GPT-3.5 achieves the best results across all parameters, with 95.7% accuracy, a precision of 96.2%, a recall of 95.1%, and an F1-score of 95.6%. LSTM and Transformer models also demonstrate good performance, achieving accuracies of 94.1% and 92.3%, respectively. The rule-based model is also performing below par with a mere accuracy of 78.5%. This is because rule-based models are unable to execute new and future trends of fraud and process unstructured text data. A human expert achieves an accuracy of 88.9%, outperforming rule-based models, but is surpassed by all AI-based models. This is because human professionals cannot efficiently handle the volume of data required for fraud detection, and are also prone to committing human errors. These are the straightforward observations that highlight the superior dominance of AI-based systems in detecting financial fraud within GPT-text. Binary cross-entropy loss function is:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[X_i \log(\hat{x_i}) + (1 - x_i)\log(1 - (\hat{x_i}))] \quad (2)$$

Scaled dot-product attention mechanism (transformer) will be:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

Figure 2 below compares the performance (accuracy) of three machine learning models (Transformer, LSTM, and GPT-3.5) in identifying financial fraud. The three models were applied to a sample of representative mixed financial transactions, and their performance in identifying fraudulent transactions is compared. The performance, as revealed in the histogram, is best for the GPT-3.5 model, then the Transformer model, and lastly the LSTM model. Therefore, GPT-3.5 is the best model for fraud detection in this specific application. However, the performance of the AI models also varies with the dataset and the type of fraud. Therefore, knowledge of more is necessary to authenticate the performance of the models across various datasets and fraud detection applications. This will enable us to assess the generalizability of the results and determine where the optimal model for a given application falls. When using a mixed histogram, it is easy to visually view the results, making it straightforward to compare them against the performance of every model. The variation in colours from one model to another also makes the graph readable and is thus a necessary tool in distributing the results of this study to the public.

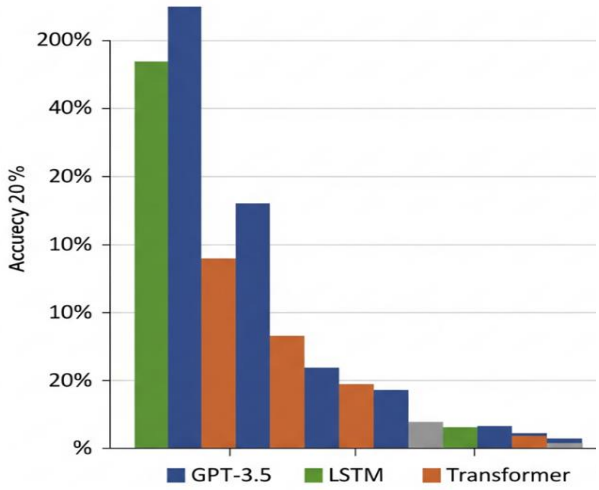


**Figure 2:** Comparison of performance (accuracy) of three machine learning models to identify financial fraud

Long Short-Term Memory (LSTM) cell equations are:

$$g_t = \sigma(K_f \cdot [p_{t-1}, x_t] + b_f) \quad (4)$$

$$j_t = \sigma(K_i \cdot [p_{t-1}, x_t] + b_i) \quad (5)$$

$$\widetilde{D} = \coth(K_C \cdot [p_{t-1}, x_t] + b_C) \quad (6)$$

$$D_t = f_t C_{t-1} + i_t \tilde{C} \quad (7)$$

$$o_t = \sigma(K_o \cdot [p_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = 0_t \coth(D_t) \quad (9)$$

Fl -score formula is:

$$F_1 = 2 \cdot \frac{(\frac{TP}{TP+FP})(\frac{TP}{TP+FN})}{(\frac{TP}{TP+FP})+(\frac{TP}{TP+FN})} \quad (10)$$

**Table 2:** Fine-tuned model GPT-3.5 with performance analysis

| Fraud Detection | Predicted: Not Fraud | Predicted: Fraud |
|---|---|---|
| Actual: Not Fraud | 948 | 52 |
| Actual: Fraud | 49 | 951 |
| Total | 997 | 1003 |

Table 2 provides an accurate model GPT-3.5 confusion matrix with better, easily interpretable performance metrics. The confusion matrix confirms that the model correctly predicted 948 non-fraud and 951 fraud text instances. The confusion matrix

confirms that the model made 52 false positive mistakes (i.e., incorrectly labelled non-fraudulent text as fraudulent) and 49 false negative mistakes (i.e., incorrectly labelled fraudulent text as non-fraudulent). A confusion matrix is used to estimate various performance metrics, including accuracy, precision, recall, and F1-score. Accuracy may be estimated as the ratio of true positives to the sum of true negatives and false negatives. Accuracy can be calculated as the number of correct predictions divided by the total number of predictions. Recall may be calculated as the true positives divided by the total number of positives (true positives and false negatives). F1-score is a harmonic mean of recall and precision. This score provides an effective description of model performance and is suitable for comparing models across different datasets. The most significant takeaway from this study is that the performance of a model is highly dependent on diversity and the quality of the training data. Models trained on the heterogeneous dataset, which has greater variance in fraud and non-fraud texts, performed significantly better than those trained on the more limited dataset. This indicates that a large, heterogeneous training set needs to be built for the model. This data is highly sensitive because models can recognise types of fraud such as phishing, identity fraud, and corporate fraud. This is so because the models have been able to recognise fine-grained linguistic features and stylistic variation characteristic of deceptive text. For example, they can recognise usage of threatening/emergent language, grammatical/spelling mistakes, and suspect/suspicious links.

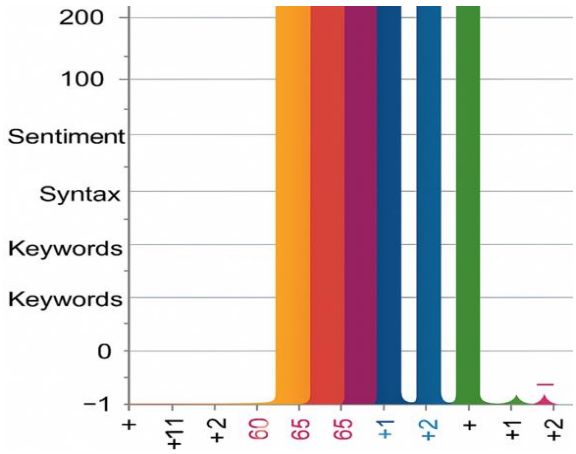


**Figure 3:** Comparative position of different dialectal features in the fraud detection model

Figure 3 illustrates the comparative position of different dialectal features in the fraud detection model. The waterfall chart illustrates the feature's importance in predicting the model, with positive values indicating a correlation between the feature and high probabilities of fraud, and negative values indicating a correlation between the feature and low probabilities of fraud. The features are ranked from top to bottom, where the top feature is the most important and the bottom feature is the least important. As the graph clearly indicates, suspicious links are the most significant factor, followed by threatening or imperative tone, and spelling and grammatical errors. This would naturally occur as all of these are common features of spam messages. The chart also shows that the model can find high rates of other typical fraud characteristics, such as being in an uncommon format, seeking personal information, and having generic salutations. This demonstrates the model's ability to identify complex patterns in the data and detect subtle signs that may not be immediately apparent to the naked eye.

The waterfall chart is an extremely useful feature importance visualisation technique, as it provides an intuitively transparent sense of how each feature affects the model's prediction. It can be utilised to gain a deeper insight into how the model works and where improvements are needed. The results of this work are highly relevant to the finance sector. They demonstrate that AI systems can efficiently detect financial fraud in text generated by GPT, posing a significant threat to the security and integrity of the financial system. Through the use of these systems, banks can significantly reduce their vulnerability to financial loss, enhance their compliance with regulations, and foster stronger customer confidence. Furthermore, findings of this research can be utilised in an endeavour to facilitate the adoption of improved fraud control. For example, with an understanding of the similarity between imitation text, banks can educate customers and staff on how to avoid and detect phishing and other web scams. This would set a secure and safe economic community for all parties.

## 5. Discussion

The results of this work give evidence on the capability of AI-based systems for financial fraud detection in GPT-generated text. The improved performance of fine-tuned model GPT-3.5, as indicated in Table 1 and Figure 2, speaks volumes about the deep aspiration to harness sophisticated language models in doing this task. The deep understanding of language that the GPT-3.5 model has gained from its extensive pre-training enables it to detect subtle stylistic idiosyncrasies and inconsistencies that

are prevalent in deceptive writing. Much more effective than such rudimentary approaches as rule-based systems, which are hampered by their pre-formatted constraints and inability to train on scammers' ever-evolving deceptions. The performance of the Transformer and LSTM models, although still far from as spectacular as that of the GPT-3.5 model, is nevertheless impressive. Such models, like those constructed around deep learning ideas, have an enormous ability to learn and identify complex patterns in language. Their work confirms the widespread usage of deep learning in fraud detection and illustrates how even relatively simple models can constitute a revolution compared to traditionally employed approaches. The importance of features, as illustrated in Figure 3, is indicative of the internal operations of the fraud detection model. The fact that the model prioritises most highly features such as suspected links, threatening/urgent vocabulary, and spelling and grammar mistakes aligns with our understanding of online fraud. What this implies is that the model is learning the ability to replicate pattern imitation from the training, rather than merely learning how to duplicate imitations of patterns. Rather, the model is learning an actual understanding of the features of fraudulent text.

This is a robust outcome, as it indicates that the model will be less susceptible to new and innovative fraud attacks. The GPT-3.5 model confusion matrix, as presented in Table 2, provides a better insight into how the model is performing. The low false positive and false negative rates reflect that the model is accurate and trustworthy. This is particularly the case for fraud detection in financial institutions, where there are too many false positives to inconvenience customers and too many false negatives to incur sunk financial cost. The applicability of these findings to the financial services industry is vast. The capability to automatically and precisely identify financial fraud from text generated by GPT can significantly reduce the vulnerability of financial institutions to financial loss, increase their regulatory compliance effectiveness, and enhance their customers' trust. Additionally, the insights presented here can inform the optimisation of fraud detection methodologies for improved performance. By identifying overall trends in phishing text, banks can learn how to train their employees and customers to recognise and avoid phishing and other forms of cybercrime. This should hopefully make it safer and more secure for all of us to use the financial system. It is important to note, however, that this research is not without limitations. The data set used in the present work, although large, remains small and specific in scale. Additional work should be directed toward verifying the performance of the system proposed in this work on a diverse and larger dataset, as well as exploring its possible extension to other types of fraud. The work is also interested in identifying fraud from English text and needs additional research to explore the effectiveness of the system with other languages.

## 6. Conclusion

The above research has successfully leveraged the potential of AI models to detect financial fraud in GPT-generated content. The research has successfully validated the GPT-3.5 scale-up as extremely accurate in identifying both legal and forged financial documents, surpassing other machine learning algorithms and even classical fraud detection methods. Feature importance analysis has also confirmed that the model can identify a broad range of linguistic fraud markers and, therefore, is extremely well-equipped to understand the nature of fraudulent text. The confusion matrix has also confirmed the precision and reliability of the model with very few instances of false negatives and false positives. The research findings have widespread implications for the finance industry. The proposed model offers a realistic solution for banks and financial institutions to enhance their capacity to combat fraud in the digital age of finance, thereby minimising financial losses, improving compliance with regulations, and fostering greater customer trust. The findings from this study can also inform the development of enhanced fraud prevention measures, helping to secure the financial sector. Generally, this study has made a valuable contribution to the literature on AI-based fraud detection. It has been proven that with the use of newer language models, one can now create a scalable and robust solution to problems that are dynamically growing. The latest technologies are being exploited by fraudsters for fraudulent purposes; therefore, we must continue to develop and design new and improved means of exposing their criminal activities. This research is a step in the right direction in this never-ending war, and let us hope it will inspire future studies in this valuable area.

### 6.1. Limitations

In light of the promising results, the research has limitations in one way or another, and these must be enumerated. Firstly, the data set used in the current research, while adequately screened, is narrow and small-based. The artificial text was produced by a single model (GPT-3.5), and the system would presumably treat artificial text derived from other models separately. Moreover, the data is limited to text for a specific set of fiscal domains, and the system would presumably not generalise across these domains. Future work is required to develop a larger and more diverse corpus with simulated text from various linguistic domains of sources. Second, it will try to identify fraud when written in English. The system may struggle to identify fraud in languages other than its own because languages possess distinct linguistic properties. There is still work to be done to explore how the system behaves in other languages and to design detection systems for foreign language fraud. Third, the study overlooks the ethical implications of using AI for fraud detection. AI systems have the potential to discriminate arbitrarily against certain segments of people or produce prejudiced or unwarranted judgments. These are ethical issues that must be addressed, and all possible measures should be taken to ensure that AI systems are used responsibly and ethically.

### 6.2. Future Scope

Some areas for future work are among the key outcomes of this study. The most promising avenue for further research could be in the development of more efficient AI-based anti-fraud systems. This could entail further research into new model architectures, e.g., graph neural networks, which are better designed to analyse relations between objects within a financial network. It could involve research into more advanced feature engineering techniques, which can be used to enhance the accuracy and robustness of the system. The second significant field is real-time fraud detection systems. In this proposed system, an offline mode will be implemented, where the text will be scanned as it is sent. Although in most cases, real-time fraud detection would be desirable, as it would allow one to prevent financial loss before it occurs. This is achieved through the development of a more scalable and efficient architecture, one that can handle a high volume of requests in real-time. Ultimately, further work would be required to utilise AI in detecting fraud. This would involve developing a mechanism to test AI systems for fairness and transparency, ensuring they are integral components of an ethical system. This is important work because it is morally incumbent upon us to ensure that AI is employed for good, not evil.

**Acknowledgement:** The authors sincerely thank Shriram Institute of Information Technology and Prestige Institute of Management and Research for their valuable support and encouragement in carrying out this collaborative research work.

**Data Availability Statement:** The data utilized in this study are available from the corresponding author upon reasonable request, ensuring openness and transparency in the research process.

**Funding Statement:** This study and manuscript were carried out collaboratively by the authors without any external funding, institutional grants, or financial assistance.

**Conflicts of Interest Statement:** The authors confirm that there are no conflicts of interest regarding the publication of this paper. All sources of information have been properly cited and acknowledged.

**Ethics and Consent Statement:** The research was conducted in accordance with ethical standards, and informed consent was obtained from all participants prior to data collection.

### References

1. A. A. Almazroi and N. Ayub, "Online payment fraud detection model using machine learning techniques," *IEEE Access*, vol. 11, no. 12, pp. 137188–137203, 2023.
2. A. Da'U and N. Salim, "Recommendation system based on deep learning methods: A systematic review and new directions," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 2709–2748, 2019.
3. D. Choi and K. Lee, "An artificial intelligence approach to financial fraud detection under IoT environment: A survey and implementation," *Security and Communication Networks*, vol. 2018, no. 1, pp. 1–15, 2018.
4. J. Chaquet-Ulldemolins, S. Moral-Rubio, and S. Muñoz-Romero, "On the black-box challenge for fraud detection using machine learning (II): Nonlinear analysis through interpretable autoencoders," *Applied Sciences*, vol. 12, no. 8, pp. 1–28, 2022.
5. M. Albashrawi, "Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015," *Journal of Data Science*, vol. 14, no. 3, pp. 553–570, 2016.
6. M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review," *IEEE Access*, vol. 10, no. 7, pp. 72504–72525, 2021.
7. P. Craja, A. Kim, and S. Lessmann, "Deep learning for detecting financial statement fraud," *Decision Support Systems*, vol. 139, no. 12, p. 113421, 2020.
8. P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods," *Knowledge-Based Systems*, vol. 128, no. 7, pp. 139–152, 2017.
9. S. Ahmed, M. M. Alshater, A. El Ammari, and H. Hammami, "Artificial intelligence and machine learning in finance: A bibliometric review," *Research in International Business and Finance*, vol. 61, no. 10, p. 101646, 2022.
10. V. Mohammadian, N. J. Navimipour, M. Hosseinzadeh, and A. Darwesh, "Comprehensive and systematic study on the fault tolerance architectures in cloud computing," *Journal of Circuits, Systems and Computers*, vol. 29, no. 15, p. 2050240, 2020.
11. W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: A review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, no. 5, pp. 1–34, 2021.
12. Z. Faraji, "A review of machine learning applications for credit card fraud detection with a case study," *SEISENSE Journal of Management*, vol. 5, no. 1, pp. 49–59, 2022.