Exploring the Boundaries of Financial Statement Fraud Detection with Large Language Models

Georgia Boskou [a], Evrikleia Chatzipetrou [a], Eleftherios Tiakas [a], Efstathios Kirkos [a], Charalampos Spathis [b]

[a] Department of Accounting and Information Systems, International Hellenic University, PO BOX:141, 57400, Sindos, Thessaloniki, Greece

[b] School of Economics, Aristotle University of Thessaloniki, N.O.P.E. Building, University Campus, 54124 Thessaloniki, Greece

Corresponding Author: Efstathios Kirkos

Corresponding Author Address: Department of Accounting and Information Systems, PO BOX:141, 57400, Sindos, Thessaloniki, Greece

Corresponding Author email: stkirk@ihu.gr

Declaration of interest: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the authors used Chat GPT 4 in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Exploring the Boundaries of Financial Statement Fraud Detection with Large Language Models

Abstract

Financial statement fraud is a significant concern in today's world, representing the most financially destructive form of occupational fraud. It negatively impacts a wide array of stakeholders, including the organization itself, its investors, creditors, suppliers, employees, and the general public. Given its critical importance, this issue has been a focal point of scholarly research over the past decades. Previous studies have utilized a diverse array of statistical, machine learning, and deep learning techniques to analyze both arithmetic and textual data. However, the use of Large Language Models (LLMs) for detecting financial statement fraud remains largely unexplored. Recent scholarly work underscores the potential of LLMs in various linguistic analyses and auditing tasks. In this study, we examine the ability of an LLM, specifically ChatGPT-4, to identify fraudulent financial statements by analyzing text segments extracted from annual reports. We focus particularly on the CEO's letters to shareholders and sections discussing the company's risk exposures. Rather than employing complex fine-tuning, we utilize prompt engineering techniques, that can be easily implemented by auditors without machine learning expertise. The findings are promising, demonstrating the potential of LLMs to uncover financial statement falsifications. By incorporating human feedback into the model's classification results, we achieved a performance score of 67% across sensitivity, specificity, and F-Measure metrics. With the ongoing growth in model size and capabilities, coupled with active research in prompt engineering, we anticipate further improvements in performance shortly. This research could prove invaluable to auditors, lenders, investors, and regulatory bodies.

Keywords: Financial Statement Fraud, Large Language Models, Prompt Engineering, Auditing

1. Introduction

Occupational fraud, which involves individuals committing fraud against their employers, is a recognized issue in contemporary society. According to the Association of Certified Fraud Examiners, the premier global anti-fraud organization, their 2020 Report to the Nations (ACFE, 2020) disclosed that occupational fraud resulted in total losses exceeding $3.6 billion from January 2018 to September 2019. The three primary types of occupational fraud identified are asset misappropriation, financial statement fraud, and corruption. Of these, financial statement fraud, while least frequent—comprising approximately 10% of cases—is the most financially damaging, with median losses reaching $954,000 (ACFE, 2020). In instances of financial statement fraud, perpetrators deliberately manipulate financial records to present a distorted view of the organization's financial health, often inflating revenues, profits, and assets, while minimizing reported losses, expenses, liabilities, and debts. The implications of financial statement fraud are far-reaching, adversely affecting a variety of stakeholders, including the organization itself, investors, creditors, suppliers, employees, and the general public

The significance of financial statement fraud has kept it under intense scrutiny for several decades. Historical financial scandals during the 1970s and 1980s, such as the manipulation of financial statements by Equity Funding Corporation through its insurance program and the fraudulent activities of ZZZZ Best, which included check kiting and insurance scams, have spurred extensive academic research. Initially, research focused primarily on corporate governance and stock market dynamics. Beasley (1996) found that companies without fraud typically had boards composed of a higher percentage of external members compared to those involved in fraud, although the presence of an audit committee did not significantly deter fraud occurrences. Summers and Sweeney (1998) observed that insiders often decrease their stock holdings in companies engaged in fraud through elevated selling activities. Further research by Beasley et al. (2000) indicated that companies engaged in fraud tended to have fewer and less independent audit committees and fewer audit committee meetings. Other researchers have investigated the use of financial ratios as indicators of fraud. Financial ratios are crucial tools in financial analysis, providing insights into a company's financial health, operational efficiency, and market performance. The premise is that anomalies in financial ratios may reflect fraudulent activities. For example, Lee et al. (1999) utilized logistic regression to demonstrate that cases of fraud typically show an excessive discrepancy between earnings and operating cash flow. Bell and Carcello (2000) integrated financial ratios with corporate governance variables in a logistic regression model to improve fraud detection capabilities. Spathis (2002) also employed financial ratios and logistic regression, analyzing published data to identify fraudulent financial statements.

Machine Learning (ML) techniques, which involve algorithms that learn from data, have been instrumental from the onset of research in this field. In their pioneering study, Green and Choi (1997) utilized endogenous financial data to develop a Neural Network model for detecting financial statement fraud, affirming the potential of Neural Networks (NNs) as tools for fraud risk assessment. Similarly, Fanning and Cogger (1998) applied Neural Networks using publicly available data to identify management fraud. A notable early contribution was made by Kirkos et al. (2007), who compared three different ML techniques—Decision Trees, Neural Networks, and Bayesian Belief Networks—using publicly available financial data to detect financial statement fraud. Their findings, which reported high accuracy levels, significantly propelled subsequent research. A considerable body of research followed, employing a variety of machine learning and data mining methodologies alongside financial, macroeconomic, and other quantitative variables to develop models capable of predicting financial statement fraud, as evidenced by works such as Ravisankar et al. (2011), Perols (2011), and Omar et al. (2017). Starting in the early 2020s, a new trend appeared in the relevant research literature. Qualitative data, i.e. textual data extracted mainly form annual reports attracted the interest of the researchers. Text segments, in combination with Text Mining and Natural Language Processing methodologies have been used for the detection of financial statement fraud (Goel et al., 2010; Humpherys et al., 2011; Purda & Scilicorn, 2014; Minhas & Hussain, 2016).

Recently, Large Language Models (LLMs) have surged in popularity. LLMs fall under the umbrella of Generative Artificial Intelligence, a category of systems which accept as input prompts (text), images, sound and video and produce as output original content of several types. Specifically, LLMs take textual prompts and generate contextually relevant and original text. The foundation of modern LLMs can be traced back to the seminal paper by Vaswani et al. (2017), which introduced the Transformer, a new type of deep neural network. The Transformer model is notable for two major innovations: positional encoding and the Self-Attention mechanism. Positional encoding tracks the position of each term within the text, facilitating the understanding of sequence order. The more critical innovation, Self-Attention, allows the model to evaluate the relevance of each word within a sequence based on its context. This mechanism assigns weights to each word, considering its relationship to every other word in the sequence, thus enabling the model to achieve a comprehensive overview of the sentence and to model relationships between words regardless of their distance in the text. This is a significant advancement over previous neural network architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which could only capture dependencies between adjacent words and struggled with modeling relationships between distantly positioned words.

In November 2022, OpenAI launched ChatGPT 3.5, a Large Language Model (LLM) that utilizes the Transformer architecture. ChatGPT quickly became the fastest-growing consumer software application in history, amassing over 100 million users within just two months (Hu, 2023). ChatGPT, like other LLMs, employs a version of the Transformer model which requires an extensive pretraining phase. During this phase, the model is trained on a vast corpus of text, often reaching the scale of terabytes. This pretraining is crucial as it constitutes the core training period for the model, necessitating significant computational resources and time. Transformers, as well as other neural networks, are fundamentally data-driven systems, meaning the quality of the training data is as critical as the architecture of the network itself. Even a sophisticated neural network can perform poorly if trained with suboptimal data, leading to issues such as bias and overfitting. There are several publicly available datasets for training LLMs, with Common Crawl being among the most prominent. Once pretrained, these models are capable of executing a variety of tasks, including summarization, translation, generating on-demand responses to queries, and even producing text and software source code.

Upon completing the pre-training stage, models achieve a remarkable grasp of language and a substantial capability to handle diverse tasks across various application domains. However, this does not equate to them possessing "expert" skills in specific fields. They are not simultaneously specialized doctors, lawyers, chemists, geologists, etc. Specialized knowledge is essential for specialized tasks. This expertise is imparted to models through a process known as Fine Tuning, which transforms general-purpose models into specialized ones. While it is technically possible to retrain a model from scratch, such an approach is costly and time-consuming. Fine Tuning leverages the broad knowledge base acquired during pre-training and builds upon it by focusing on additional, more focused capabilities. This involves training the model further with a smaller, specialized dataset. Though less resource-intensive than pre-training, Fine Tuning still demands a dataset of adequate size and quality, sufficient hardware resources, a deep understanding of ML principles, familiarity with the model's architecture, and programming skills.

Interestingly, Fine Tuning is not the only method for influencing the output of a model. Models are highly responsive to the prompts they receive, meaning that variations in prompts on the same topic can lead to different outcomes. Through prompts, users can define a role for the model, issue instructions, provide relevant examples for learning, and specify the desired form of the output. Andrej Karpathy, a founding member of OpenAI, made a compelling statement on Twitter, declaring, "The hottest new programming language is English." He elaborated that LLMs are capable of performing in-context learning and can be programmed using input-output examples. The term "Prompt Engineering" refers to the process of developing effective techniques for writing prompts that enhance the model's performance on specific tasks. Systematic research in the field of prompt engineering has produced notable results. Researchers claim that by practicing prompt engineering have achieved comparable or better results to other systems created with specialized fine tuning (Nori et al., 2023; Gunawardena et al., 2023).

In the present study we explore the capability of a LLM, specifically ChatGPT 4, to detect fraudulent financial statements through the analysis of text segments extracted from annual reports. Former research findings suggest that it is feasible to distinguish fraudulent financial statements from legitimate ones by conducting contextual and linguistic analyses on narrative sections of annual reports (Goel et al., 2010; Humpherys et al., 2011; Glancy & Yadav, 2011; Minhas & Hussain, 2016; Goel & Uzuner, 2016). LLMs, with their extensive knowledge of language patterns and structures, are well-suited for classification tasks such as sentiment analysis (Vasarhelyi et al., 2023; Gu et al., 2023; Hillebrand et al., 2023). Additionally, by utilizing prompt engineering techniques like instructional learning and in-context learning, LLMs can be guided to acquire specific knowledge and optimize their output for specialized tasks. Our objective is to explore the inherent capabilities of LLMs for detecting fraudulent financial statements, leveraging prompt engineering but without the application of fine tuning. While fine tuning can enhance the capabilities of an LLM, it demands expert machine learning knowledge and the creation of a meticulous training set, a tedious and time consuming task.

The structure of the paper is as follows: Section 2 provides a review of the literature in this field, Section 3 outlines the methodology used, Section 4 discusses the findings, and Section 5 offers conclusions.

2. Literature Review

Given the significant ramifications of financial statement fraud, this topic has been extensively explored by numerous researchers. Early pioneers in applying machine learning (ML) techniques to detect financial statement fraud include Green and Choi (1997) and Fanning and Cogger (1998), who were among the first to utilize neural networks for this purpose. Several notable following studies have gained broad recognition within the academic community: Dechow et al. (2011) conducted a detailed analysis of 2,190 cases sourced from the SEC's Accounting and Audit Enforcement Releases. They employed a variety of variables, provided insights into their relevance, and introduced the F-Score as a metric to identify fraudulent cases. Kirkos et al. (2007) examined 76 Greek instances of financial statement manipulation, analyzing publicly available financial ratios using Decision Trees, Neural Networks, and Bayesian Belief Networks. Their models not only achieved high accuracy rates but also identified the most influential variables. Beneish (1999) demonstrated a systematic relationship between the likelihood of manipulation and certain financial statement variables using the Weighted Exogenous Sample Maximum Likelihood (WESML) probit method. Ravisankar et al. (2011) assessed the efficacy of various data mining techniques, including Neural Networks, Support Vector Machines, Genetic Programming, Group Method of Data Handling, and Logistic Regression, in fraud detection. Their research also included experiments with feature selection, further refining the approach to identifying fraudulent activities.

Until the early 2020s, the focus of research predominantly centered on the analysis of numerical-quantitative data. However, Goel et al. (2010) explored the verbal content and presentation style of the qualitative segments of annual reports, concluding that linguistic features could effectively detect fraud. Cecchini et al. (2010) developed dictionaries from 10-K filings to predict fraud and bankruptcy, enhancing model performance by integrating quantitative and qualitative data in the input vectors. Glancy and Yadav (2011) introduced a Computational Fraud Detection Model that applied a quantitative approach to textual data. Humpherys et al. (2011) critiqued earlier studies for their narrow focus on numerical data, overlooking the explanatory text accompanying financial statements. They pointed out that the Accounting and Auditing Enforcement Releases (AAERs) associated with fraudulent financial reports often revealed deceptive communication, misdirection, and obfuscation within the text-based portions, suggesting that these segments are valuable for fraud investigation. The authors employed linguistic constructs and variables such as affect, complexity, and uncertainty, along with machine learning techniques. Although the accuracies of these models were relatively modest, ranging between 60% and 67%, the study is considered a foundational step for subsequent research. Subsequent studies have continued to enrich the scientific understanding of using textual-qualitative data for detecting financial statement fraud, as evidenced by the contributions of Goel and Uzuner (2016), Purda and Scilicorn (2014), and Minhas and Hussain (2016). Hajek and Henriques (2017) provided a comparative study that mines corporate annual reports for intelligent detection of financial statement fraud, incorporating analyses of both quantitative and qualitative data.

Numerous studies have conducted comprehensive reviews of the literature related to financial statement fraud. Shahana et al. (2023) analyzed 84 papers to identify leading authors and to outline the most commonly used fraud detection techniques, performance metrics, datasets, and class balancing methods. Soltani et al. (2023) combined bibliometric analysis techniques with topic modeling to develop a framework that integrates machine learning into the analysis of financial statement fraud literature. Gupta and Mehta (2021) performed a Systematic Literature Review, examining articles published from 1995 to 2020.

Ashtiani and Raahemi (2022) critically synthesized and analyzed 47 papers on financial statement fraud in their systematic literature review. In what follows, we will review the most recent articles, published between 2021 and 2023.

The detection of fraudulent financial statements continues to be a vibrant area of research today. Notably, a substantial proportion of recent studies primarily employ machine learning and advanced deep learning techniques, demanding a thorough understanding of these methods as well as programming skills. For instance, Wang et al. (2023) introduced the RCMA model, which incorporates a Ratio-aware, Chapter-aware, and Modality-aware Attention mechanism. The Ratio-aware Attention mechanism utilizes a Multilayer Perceptron to analyze financial ratios, while the Chapter-aware mechanism employs a Long Short-Term Memory (LSTM) network. The Modality-aware attention layer synchronizes the capabilities of the different modalities. Li et al. (2024) conducted a detailed analysis of the Management Discussion & Analysis (MD&A) section, focusing on three dimensions: textual language structure, language quality, and language expression. They identified six textual indicators and combined them with eleven financial variables to create the input vector. Common machine learning methods such as Multilayer Perceptron Neural Networks, XGBoost, and Logistic Regression are employed for the classification tasks.

Achakzai & Peng (2023) explore the use of both financial and non-financial variables as inputs and introduce the Dynamic Ensemble Selection algorithm, which dynamically combines individual classifiers for improved performance. Zhang et al. (2022) delve into the MD&A segment of annual reports, utilizing Word2Vec for word encoding and applying machine learning techniques such as Random Forest, SVM, and Naïve Bayes to detect fraudulent statements. Cheng et al. (2021) tackle several prevalent issues in data science, including class imbalance, missing values, and feature selection. Their approach involves using purely quantitative data and typical machine learning methods. Riskiyadi (2023) also focuses on quantitative data, incorporating both financial and non-financial ratios and employing well-known machine learning algorithms to develop predictive models. Guan et al. (2022) combine quantitative data with Logistic Regression to enhance their analysis framework. Yadav and Sora (2022) adopt a hybrid approach using a Convolutional Neural Network that integrates textual features with financial variables. They contribute significantly to the field by developing an innovative algorithm for feature selection, clustering and labeling the selected features, and optimizing the hyperparameters of the CNN using the MRFO learning algorithm.

Purda and Skillicorn (2014) conducted an analysis of the language employed in the Management Discussion & Analysis (MD&A) section of quarterly and annual reports. They utilized a Bag of Words approach and applied the Random Forest algorithm to rank words from most to least predictive. Based on the top 200 words identified, they trained a Support Vector Machine classifier to enhance detection capabilities. Huang et al. (2022) addressed the issue of misclassification costs associated with false positive and false negative predictions by proposing a cost-sensitive cascade forest tailored for fraud detection. They also examined the impact of various missing value imputation methods on their analysis. The data employed in their study consisted of financial ratios, providing a quantitative foundation for their models.

The application of Large Language Models (LLMs) in accounting and auditing is increasingly capturing the interest of researchers. The intricate understanding of language structures and patterns, coupled with their reasoning abilities, inherent knowledge, the ability to communicate in natural language with humans and to modify their output according to the given prompts, render LLMs invaluable tools for accountants and auditors. These professionals, who may not be experts in machine learning, can now seamlessly interact with advanced artificial intelligence systems to perform complex tasks that require expert knowledge and judgment. Recent studies highlight the significant benefits and potential of LLMs in the fields of accounting and auditing. Vasarhelyi et al. (2023), in a commentary article, discuss the transformative potential of LLMs in Accounting. The paper addresses issues related to accounting education, research and professional auditing. Gu et al. (2023) introduce the concept of co-piloted auditing, where the model is not just a tool used by the auditor, but it operates synergistically with humans. The authors explain fine tuning and prompt engineering, propose tasks for the LLM through all phases of the audit procedure, define an audit prompt protocol and evaluate prompting strategies applied for financial ratio analysis, post-implementation review and journal entry testing. Li et al. (2024) assesses the ability of a LLM to cross-verify internal accounting records using external textual evidence in a continuous audit environment and demonstrate its superiority over manual approaches. Fotoh and Mugwira (2023) claim that external auditors can benefit from ChatGPT in five key areas including the analysis of financial data, the processing of large volumes of unstructured data, the extraction of information relevant to auditors, risk assessment and risk projections and the automation of audit process. The authors also discuss several limitations and ethical considerations. Hillerbrand et al. (2023) propose the ZeroShotALI, a novel recommender system which links paragraphs in a financial document to their corresponding regulatory requirements of IFRS which are presented as a collection of checklist items. The system helps auditors to identify relevant text segments for each checklist item.

However, the application of Large Language Models (LLMs) for detecting management fraud remains a largely unexplored area. Bhattacharya and Mickovic (2024) have advanced this field by fine-tuning the BERT LLM using the Management Discussion and Analysis section from 10-K annual reports to create an ensemble model that integrates quantitative and textual data. They enhanced the model's architecture by incorporating a sigmoid layer at the end of the final layer to assess the likelihood of fraud. The modified model, termed $\text{Bert}_{\text{final}}$, demonstrated high performance. However, their methodology comes with significant limitations due to the advanced expertise and resources it demands. Fine-tuning such models requires a profound understanding of the model's architecture, adeptness at adjusting crucial hyperparameters (such as learning rate and epochs), specific programming skills, and access to suitable hardware. Additionally, the process involves the laborious task of constructing a training and validation dataset—Bhattacharya and Mickovic (2024), for instance, compiled data from 30,876 firm-year observations. These requirements make the approach challenging to implement without substantial expertise in machine learning.

The examination of the aforementioned studies reveals that detecting financial statement fraud remains a vibrant area of research today. Researchers commonly use machine learning and occasionally deep learning techniques, analyzing both numerical-quantitative and textual-qualitative data. To date, the use of Large Language Models (LLMs) in this context has been explored exclusively by Bhattacharya and Mickovic (2024), who modified the architecture of an LLM and conducted fine-tuning. To the best of our knowledge, no other study has yet investigated the applicability of LLMs in detecting financial statement fraud by relying solely on the model's inherent knowledge, enhanced through prompt engineering techniques. This study aims to make a contribution in this direction. Our research question is defined as follows: Can ChatGPT-4 detect fraud in financial statements by analyzing text segments from annual reports, utilizing its inherent knowledge augmented with prompt engineering guidance and training?

3. Methodology

In this study, our goal is to assess the capability of ChatGPT to detect cases of financial statement fraud by exclusively employing prompt engineering, without any fine-tuning. This approach treats ChatGPT as an end-to-end tool for fraud detection. LLMs are efficient in several Natural Language Processing (NLP) tasks. One could use an LLM as an NLP tool (eg. as a Part of Speech Tagging tool, or for calculating the length of the sentences (as a mean of text complexity)) to extract linguistic features. Then one could use these features as predictors in another platform. For example, Hillebrand et al. (2023) utilizes an LLM as a text retrieval tool and subsequently compares the retrieved text with other texts using cosine similarity. While some researchers modify the original architecture of an LLM and apply fine-tuning to enhance fraud detection capabilities, such as Bhattacharya and Mickovic (2024) who added a sigmoid layer to the end of a BERT model, this study focuses on a different aspect. The central question here is whether auditors—experts in their field, but not in machine learning—can effectively use an LLM through natural language to aid in the identification of fraud in annual reports. The tool of choice for this task is prompt engineering.

Prompt Engineering

LLMs are pretrained using a self supervised approach. This entails feeding the models vast amounts of text data to facilitate their learning processes. The model reads a sequence of words and then tries to predict the next most probable word. A wrong prediction causes the adjustment of the values of the model's trainable parameters. This way the model gains knowledge about linguistic structures. This type of training paradigm was the main training practice for large models until 2017. During 2017-2019 the training practice shifted to a "pretrain and fine tune paradigm". Initially the model undergoes pretraining in order to gain generic knowledge and skills. In a following stage the model is trained again by using a smaller set of task-specific data in order to gain specialized knowledge and skills. During this stage only the parameters of the last layers, or the parameters of newly added layers are tuned. Nowadays the training practice shifts to a "pretrain, prompt and predict" paradigm (Liu et al., 2021). In this paradigm specialized prompts are given to the model in order to leverage the existing knowledge and elicit proper responds, suitable for fulfilling specialized downstream tasks. The term "Prompt Engineering" refers to the practice of crafting effective prompts in order to successfully interact with an AI system and guide it to produce correct and desired responses.

Recent research finding highlight the effectiveness of prompt engineering. Researchers claim that by practicing prompt engineering have achieved comparable or better results to other systems created with specialized fine tuning (Brown et al., 2020; Wei et al., 2022; Nori et al., 2023; Gunawardena et al., 2023). As AI technologies become more integrated into various fields such as customer service, content creation, and software development, effective prompt engineering becomes a valuable skill. Consequently, prompt engineering has become nowadays a fruitful research area and many recent studies propose new and elaborated prompting techniques. Below we present some widespread prompting techniques, most of which have been used in the present paper.

A good practice involves providing clear and concise **instructions** to the model on how to execute a task. Another effective technique is **role assignment**, or role prompting, in prompt engineering. This strategy entails assigning a specific role or persona to an AI model within the prompt to guide its responses to align with the expected behaviors of that role. This approach is particularly effective because it leverages the AI's ability to adopt various perspectives or levels of expertise, thereby enhancing the relevance and appropriateness of its outputs, as noted by Ramlochan (2023).

One of the most recognized prompting techniques is **few-shot learning** (Brown et al., 2020). The term few-shot learning describes the practice of giving a small number of examples of the downstream task to the model through prompts. Alternative practices are one-shot learning, where a single example is given, and zero-shot learning where only natural language instructions but no examples are given. Brown et al. (2020) suggest that LLMs are meta-learners and that slow pretraining can be combined with fast "in-context" learning, implemented within the context activation of the model. The authors also claim that "the few-shot setting is sometimes competitive with or even occasionally surpasses state-of-the-art (despite state-of-the-art being held by fine-tuned models)". Their research findings suggest that differences in performance among zero, one and few-shot learning increase by the size of the model (i.e. the number of trainable parameters).

Another widely recognized prompting technique is **Chain of Thoughts** (CoT) (Wei et al., 2022). The idea is to decompose a complex task into intermediate reasoning steps which are called Chain of Thoughts. Afterwards, a small number of examples (few-shot learning) are fed as prompts to the model. Each example is constructed as input – chain of thoughts – output. The authors test the effectiveness of their method by using five different LLMs, they compare the results with those obtained with standard few-shot learning, and reach three principal conclusions: a) CoT is an emergent ability related to the size of the model. It has no significant effect on small models, but improves the performance of models with more than 100 billion trainable parameters, b) CoT has more beneficial effects when dealing with more complicated problems, c) CoT applied to large models compares favorably to fine-tuned, task-specific models, trained with labeled data.

Wei et al. (2023) demonstrated that the reasoning abilities and outcomes of a Large Language Model (LLM) can be enhanced by guiding the model to think in a stepwise manner, achieved by providing it with step-by-step reasoning examples. However, Kojima et al. (2022) critiqued this method for being case-specific and dependent on the human-engineered, multi-step reasoning prompts. As an alternative, they introduced the Zero-shot Chain of Thoughts (**Zero-shot-CoT**) prompting technique, which is remarkably simple. This technique merely involves appending the phrase "Let's think step-by-step" at the end of the prompt. The authors claim that "despite the simplicity, the Zero-shot-CoT successfully generates a plausible reasoning path in a zero-shot manner and reaches the correct answer in a problem where the standard zero-shot approach fails. Importantly, the Zero-shot-CoT is versatile and task-agnostic, unlike most prior task-specific prompt engineering in the form of examples". They report significantly better performance with Zero-shot-CoT compared to standard Zero-shot and note that its effectiveness increases with the size of the model.

The research on prompt engineering is very intensive and many alternative prompting techniques have been proposed. These include Self Consistency (Wang et al. 2022), Tree of Thoughts (Yao et al., 2023; Long, 2023), ART (Automatic multi-step Reasoning and tool-use) (Paranjape et al., 2023) and others.

It is uncommon for a model to provide correct and complete answers on the initial prompt. Typically, users must submit several iterations, each time scrutinizing the results and striving for improvements. Furthermore, the process of crafting a prompt involves several distinct stages. Willey et al. (2023) introduced the concept of the Prompt Development Life Cycle (PDLC). Similarly with the Software Development Life Cycle, PDLC defines stages where the users have to analyze and plan the prompt. Initially, users assess the problem, pinpoint its unique challenges, and evaluate the capabilities of the AI system. During the design stage, users formulate the appropriate query and determine which prompting techniques to employ and how to integrate them. Subsequently, users execute the prompt, review the responses, and iterate the process until they achieve a satisfactory outcome. Manually crafting prompts demands specific expertise and considerable time. Consequently, recent studies have suggested methods for automated prompt generation (Shin et al., 2020; Gao et al., 2021; Chen et al., 2024).

Data

In this research, we aimed to rely solely on the capabilities of the Large Language Model (LLM). Initially, we requested ChatGPT to identify companies implicated in financial statement falsification within the years 2009 to 2018. As a response we received a list of nine companies. Aware of the LLM's tendency for generating plausible yet incorrect answers—often referred to as "hallucinations"—we verified each case to ensure accuracy. Indeed, the list provided by ChatGPT included verifiable instances of financial statement falsification. Subsequently, we obtained the annual reports for these companies for the years during which the falsifications occurred. We also tasked ChatGPT with identifying comparable companies from the same industry sectors that had no history of financial fraud, and verified these companies' statuses before retrieving their respective annual reports for the same fiscal years. This approach helped to control for macroeconomic factors. The list of the selected companies and the corresponding years are presented in Appendix A. From these annual reports, we extracted the CEO's letters to shareholders and sections detailing the companies' risk exposures—areas most susceptible to falsification, as noted by Minhas and Hussain (2016). These sections were converted into plain text to facilitate processing by the model. Furthermore, to prevent any pre-existing model knowledge from influencing the outcomes, we replaced the names of the companies with the placeholder "COMPANY_NAME". The prepared documents will serve not only for few-shot learning but also for validating the model's proficiency in classifying unseen cases as either 'fraud' or 'non-fraud.'

Performance metrics

For the assessment of the model's performance, several well known performance metrics are used. Conventionally we characterize the fraud cases as "positive" and the non-fraud cases as "negative". Following this convention the following measures are defined:

True Positive (TP) = The number of fraud cases which have been correctly identified by the model.

True Negative (TN) = The number of non-fraud cases which have been correctly identified by the model.

False Positive (FP) = The number of non-fraud cases which have been wrongly classified by the model as fraud cases.

False Negative (FN) = The number of fraud cases which have been wrongly classified by the model as non-fraud cases.

The above mentioned measures are used for the definition of the following performance metrics:

Accuracy is the rate of correct predictions to the total number of cases.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Specificity is the rate of correct negative predictions to the total negative cases.

$$specificity = \frac{TN}{TN + FP}$$

Sensitivity (or Recall) is the rate of correct positive predictions to the total positive cases.

$$sensitivity = \frac{TP}{TP + FN}$$

Precision is the rate of correct positive predictions to the total number (both correct and incorrect) of positive predictions

$$precision = \frac{TP}{TP + FP}$$

F-Measure is the harmonic mean of sensitivity and precision.

$$F - Measure = \frac{2 * sensitivity * precision}{sensitivity + precision}$$

The abovementioned metrics are used for the assessment of the model's performance.

4. Results and discussion

In our experiments, we selected three legitimate and three fraudulent cases from a total of eighteen to serve as training data in a few-shot learning paradigm. The remaining twelve cases were reserved for the testing set, which we will use to assess the model's ability to accurately predict the correct class.

The initial experiment aimed to determine whether ChatGPT already possesses knowledge about the linguistic differences between deceptive and legitimate text segments in annual reports. We did not provide ChatGPT with any specific instructions or learning examples. Instead, we assigned it the role of an experienced auditor and requested it to identify known linguistic differences between texts authored by honest and dishonest managers. The details of the corresponding prompt are presented in Table 1.

Prompt 01: You are an experienced external auditor. Your task is to identify management fraud cases, where managers manipulated the financial results. Former research findings suggest that it is possible to classify cases as fraud or non-fraud by analyzing text segments found in the annual reports. This is also your task. I will ask to classify texts as fraud or non fraud cases. Reply if you understand your task and do nothing more
Prompt 02: According to your existing knowledge, what are the differences in the language between managers who intentionally try to deceive the shareholders and those who honestly report on the company's true performance.

Table 1. Prompts asking of existing knowledge

LLM's response was concrete and informative. It mentioned forensic accounting, behavioral finance and computational linguistics, although not referencing any specific papers, and it listed the following key differences:

- **Complexity**: Fraudulent communications often use more complex language to obscure reality. This might include longer sentences, more complex words, and a higher use of jargon. The complexity is intended to confuse or overwhelm the reader, making it harder to question the veracity of the information presented.
- **Clarity and Concreteness**: Honest reporting tends to be more direct, clear, and concrete. Managers who are not trying to deceive are more likely to provide specific details, quantifiable information, and straightforward explanations. In contrast, deceptive texts may be more abstract, with fewer details and more ambiguous terms that evade specific commitments.

- **Positive Emotion**: Deceptive texts may use overly positive language to create a misleadingly optimistic impression of the company's performance. This can include excessive use of superlatives, strong adjectives, and a focus on positive outcomes while minimizing or ignoring negative aspects.
- **First-person Pronouns**: Studies have shown varying results regarding the use of first-person pronouns. Some findings suggest that deceptive texts may have a lower frequency of first-person pronouns (e.g., "I", "we") as a way to distance the author from the deception. Others suggest an increase in their usage as a means to appear more persuasive and trustworthy.
- **Passive Voice**: The use of passive voice can be a strategy to obscure responsibility or to make actions seem less directly connected to the company or its management. Fraudulent reports might use passive constructions more frequently to avoid assigning clear responsibility for negative outcomes.
- **Uncertainty and Tentativeness**: Language that expresses uncertainty or tentativeness (e.g., "might", "could", "possibly") can be more prevalent in fraudulent reports, as it provides a way to hedge against specific claims that could be proven false. However, some honest managers might also use cautious language in uncertain times, so this characteristic must be interpreted within context.
- **Temporal Orientation**: Deceptive communication may have a different temporal focus, either by excessively discussing past achievements to distract from current problems or by focusing overly on future prospects without a solid basis for such projections.
- **Change in Language Style**: Sudden changes in language style or tone from one report to another without a clear reason can also be a red flag. This might include shifts in complexity, use of jargon, or emotional tone that seem intended to manipulate the reader's perception.

According to the response, ChatGPT appears to possess some understanding of the linguistic differences between fraudulent and legitimate texts. To determine whether ChatGPT could effectively apply this knowledge to classify text segments accurately, we provided it with twelve holdout text segments, each containing the manager's letter to the shareholders. We did not employ any prompt engineering techniques and simply asked it to classify the segments as fraudulent or non-fraudulent. The results were disappointing. ChatGPT classified all cases as non-fraudulent, failing to correctly identify a single instance of fraud. The corresponding performance scores are presented in Table 4.

In the second experiment, we employed a range of prompt engineering techniques. Specifically, we provided instructions by presenting the results of two research papers. We also incorporated a list of fraud indicators identified in other research studies. Additionally, we engaged in few-shot learning, inputting six texts from CEOs' letters to shareholders extracted from annual reports, half of which were fraud cases. For each text, we informed the model whether it was a fraud or a non-fraud case, asking it to learn from these examples to understand the context and linguistic characteristics of both types of cases. We instructed the model to internalize the differences as rules. Furthermore, we implemented zero-shot chain of thought (CoT) reasoning by appending the phrase "let's think step by step" to the end of the prompts. Following this training, we asked ChatGPT to report how the language used in fraudulent cases differs from that in non-fraudulent cases.

The prompts are presented in Table 2

Prompt 01: You are an experienced external auditor. Your task is to identify management fraud cases, where managers manipulated the financial results. Former research findings suggest that it is possible to classify cases as fraud or non-fraud by analyzing text segments found in the annual reports. This is also your task. I will give you instructions and examples to learn from. Later I will ask to classify texts as fraud or non fraud cases. Reply if you understand your task and do nothing more.
Prompt 02: Instruction Learning: [Former research addressed the issue of detecting management fraud by using narratives from annual reports. Glancy and Yadav (2011) in their paper titled “A computational model for financial reporting fraud detection” and published in the journal “Decision Support Systems”, analyze the Management Discussion and Analysis section, found in 10Ks. The idea is that the writer of the MDA was under some kind of stress that affected the writing. They conclude that it is possible to detect financial reporting fraud from the text of annual filings with the Security and Exchange Commission. Another study on the same topic is that of Minhas and Hussain (2016). The title of the paper was “From Spin to Swindle: Identifying Falsification in Financial Text” and it was published in the “Cognitive Computation” Journal. The authors analyze narratives from 10Ks. According to the authors our choice of words can reveal our inner intentions. The research question addressed rests on the premise that language deployed by truth-tellers and liars is distinct and can be distilled. The authors use the Coh–Metrix tool. They also use the dictionary of Loughran and McDonald. According to the authors there is a difference in the syntactic structure between fraud and non-fraud firms. This correlates with the view that deception in text is manifested by dense syntactic structure to reduce readability and comprehension. There is also a difference in the use of adverbs and adjectives, this can qualify the meaning of statements. Further, there is a difference in the use of connectives which can again lead to poor cohesion if used sparingly. Referential cohesion measures are also showing up as discriminators. This could again be the case that fraud firms are attempting to obfuscate the narratives through poor co-referencing. The literature indicates more negativity in fraud reports, greater use of modal words and passive verbs Some other research findings are the following A. Word quantity: Could be higher or lower in deceptive text. Generally, higher quantities of verbs, nouns, modifiers and group references B. Pronoun use: First person singular pronouns less frequent, greater use of third person pronouns. This is known as distancing strategies (reducing ownership of a statement) C. Emotion words: Slightly more negativity, greater emotional expressiveness D. Markers of cognitive complexity: Fewer exclusive terms (e.g. but, except), negations (e.g. no, never) and causation words (e.g. because, effect) and motion verbs—all require a deceiver to be more specific and precise. Repetitive phrasing and less diverse language is more marked in the language of liars. Also, more mention of cognitive operations such as thinking, admitting, hoping E. Modal verbs: Verbs such as would, should and could lower the level of commitment to facts F. Verbal nonimmediacy: ‘Any indication through lexical choices, syntax and phraseology of separation, nonidentity, attenuation of directness, or change in the intensity of interaction between the communicator and his referents. Results in the use of more informal, nonimmediate language G. Uncertainty: Impenetrable sentence structures (syntactic ambiguity) or use of evasive and ambiguous language that introduces uncertainty (semantic ambiguity).

Modifiers, modal verbs (e.g. should, could) and generalizing or ‘‘allness’’ terms (e.g. ‘‘everybody’’) increases uncertainty H. Half-truths and equivocations: Increased inclusion of adjectives and adverbs that qualify the meaning in statements. Sentences less cohesive and coherent thereby reducing readability I. Passive voice: Increase in use, another distancing strategy—switch subject/object around J. Relevance manipulations”: Irrelevant details K. Sense-based words: Increase use of words such as see, touch, listen] Learn from these instructions When you are ready answer “I am ready” and do nothing more.
Prompt 03: Few Shot Learning: I will give you six texts. Each text comes from the annual report of a company and includes the CEO’s letter to the shareholders. The three texts are from annual reports of companies whose management has committed fraud. The other three texts are from annual reports of companies whose management has not committed fraud. Your task is to use your own language knowledge, to use the instructions provided in prompt 02, to analyze the examples provided, and learn the context, language characteristics and structures which are common in the fraud cases, the context, language characteristics and structures which are common in the non fraud cases, and to detect differences in context, languages characteristics and structures between the fraud companies and the non fraud companies. Remember these differences as RULES which differentiate the fraud from the non fraud cases.
Prompt 04-09: This is the (first, second etc.) company. This company is a (non) fraud case. The text to analyze is between brackets. The name of the company has been substituted for COMPANY_NAME. […] Let’s think step by step. Learn the context, the language structures and characteristics of the (non) fraud case. Use them as rules. When you are ready answer “I am ready” and do nothing more.
Prompt 10: Explain how the language of the fraud cases is different from the language of the non fraud cases.

Table 2. Prompts with instruction learning and few shot learning

ChatGPT responded by providing explanations about differences in the context and language structures of the fraudulent and non-fraudulent cases. The answer of the model is presented in Table 3.

The language used in fraud and non-fraud cases, based on the examples provided, exhibits distinct characteristics that align with previous research on deceptive language and management fraud. Here’s a breakdown of the differences: **Language in Fraud Cases** 1. **Complexity and Ambiguity**: The fraud cases tend to use more complex sentence structures and ambiguous language. This complexity can obfuscate the truth, making it harder for readers to understand the real situation, consistent with findings that deceptive language aims to reduce readability and comprehension. 2. **Overly Optimistic and Vague Forward-looking Statements**: Fraud cases often contain overly optimistic projections and goals without providing concrete strategies or acknowledging potential risks in a balanced way. This aligns with the use of positive emotion words and avoidance of negative language, despite underlying issues. 3. **Use of Passive Voice**: The use of passive voice is more prevalent, serving as a distancing technique to avoid direct attribution of actions and responsibilities. This

makes it harder to pinpoint accountability.

4. **Generalizing and Abstract Language**: There's a tendency to use generalizations and abstract terms that don't commit to specific actions or outcomes. This could relate to the strategic use of uncertainty to avoid pinpointing fraud.
5. **Frequent Mention of Non-specific Innovations and Improvements**: References to innovations, improvements, or advancements are often not tied to specific, measurable outcomes, which can serve to create a facade of progress or success.

Language in Non-Fraud Cases

1. **Specificity and Clarity**: The non-fraud cases tend to provide specific information about strategies, actions taken, and results achieved. The language is more direct and clearer, making it easier for readers to grasp the company's situation and future plans.
2. **Acknowledgment of Challenges and Realistic Outlooks**: These reports are more balanced in acknowledging challenges and providing realistic outlooks. They tend to mention specific obstacles faced during the period and concrete steps taken in response.
3. **Use of Active Voice**: Active voice is more commonly used, which directly links actions to the company or management, providing a clearer picture of responsibility and agency.
4. **Detailed Descriptions of Strategies and Operations**: There's a focus on detailing strategies, operational changes, and specific areas of growth or improvement, providing transparency and enhancing credibility.
5. **Measured Optimism Tied to Specific Initiatives**: Optimism about the future is generally tied to specific initiatives, investments, or strategies, providing a foundation for positive statements and making them more credible.

Conclusion

The differences between fraud and non-fraud cases in language use are subtle yet significant. Fraud cases tend to employ strategies to obfuscate, deflect, and impress without substance, using complex language, passive constructions, and non-specific forward-looking statements. Non-fraud cases, conversely, focus on clarity, specificity, and balance, offering a more transparent and accountable narrative of their operations, challenges, and strategies. These distinctions are crucial for auditors, investors, and regulators in assessing the credibility of corporate disclosures.

Table 3. Differences in the language after instruction learning and few-shot learning.

The prompt engineering techniques we utilized significantly altered the model's responses. Initially, without prompt engineering, the language characteristics identified by the model were general and included elements such as positive emotions, first-person pronouns, and temporal orientation, which could apply to any deceptive text. In contrast, after applying prompt engineering, the context and language characteristics became more business-oriented, incorporating discussions on projections and strategies, specific initiatives, investments, actions, and outcomes. To evaluate the impact of prompt engineering on the model's ability to distinguish between fraudulent and non-fraudulent cases, we input the twelve validation texts and asked the model to classify them. The results were notably improved. The model successfully identified three out of six fraudulent cases and five out of six non-fraudulent cases. The corresponding performance scores are detailed in Table 4. The entire process is illustrated in Figure 1 as a sequence of prompts.

Minhas and Hussain (2016) posit that falsifications are also likely to occur in text segments that discuss the risks faced by a company. To enhance the predictive capabilities of our model, we enriched the input by including segments relevant to risk management and repeated the experiment. The input texts now comprised both the CEO's letter to shareholders and the sections on risk, structured into two distinct parts. The first part, labeled "CEO to shareholders," enclosed the CEO's letter within << and >> symbols. The second part, labeled "Risk management," similarly enclosed the risk-related segment. The prompts were identical to those used in the second experiment, except for the third prompt, where we clarified the new structure of the input text. Unfortunately, enriching the input text had a detrimental effect on the results. ChatGPT managed to correctly classify only one fraudulent case. However, it successfully classified all the non-fraudulent cases. This suggests that the risk-related segment did not provide useful context or language patterns for fraud identification. Furthermore, adding irrelevant text to the input appeared to confuse the model and degrade its performance. The performance scores are detailed in Table 4.

We conducted the experiment a fourth time, maintaining the settings from the second experiment. The text analyzed remained the CEO's letter to the shareholders, and the prompts were unchanged. The only modification was that after each classification result, we informed the model whether its prediction was correct or incorrect. If incorrect, we instructed the model to learn from its mistake with the prompt, "WRONG! THIS IS A (NON) FRAUD CASE. LEARN FROM YOUR MISTAKE." This approach provided human feedback, effectively turning the test stage into an additional training phase. Interestingly, the model adjusted its predictions, particularly when processing the last texts, indicating that it was learning from the feedback and modifying its behavior accordingly. Ultimately, the model correctly identified four out of six fraudulent cases and four out of six non-fraudulent cases. Comparing these results with those from the second experiment, we observed an increase in the model's ability to detect fraudulent cases, albeit at the expense of identifying non-fraudulent ones. Given that the cost of misclassifying fraud cases is typically higher, this shift suggests an overall improvement in the model's performance due to human feedback. The performance scores are detailed in Table 4.

Settings	Accuracy	Specificity	Sensitivity	Precision	F-Measure
CEO's Letter. No instruction learning, no few shot learning, no zero-shot CoT	0,5	1	0	n/a	n/a
CEO's Letter. Instruction learning + few shot learning, + zero-shot CoT	0,67	0,83	0,5	0,75	0,6
CEO's Letter + risk. Instruction learning + few shot learning, + zero-shot	0,58	1	0,17	1	0,29

CoT					
CEO’s Letter. Instruction learning + few shot learning, + zero-shot CoT + human feedback	0,67	0,67	0,67	0,67	0,67

Table 4 Performance scores.

5. Ablation study

To evaluate the impact of each prompt engineering technique, we conducted an ablation study by repeating the second experiment twice with modifications. In the first iteration, we omitted the few-shot learning component, providing only instructions without examples for learning. In the second iteration, we removed the instructional component, offering only examples without explicit instructions. The removal of few-shot learning resulted in catastrophic outcomes; the model failed to identify even a single fraudulent case, classifying all cases as non-fraudulent. These results mirrored those of the initial experiment, where no prompt engineering techniques had been applied, underscoring the significant and beneficial impact of few-shot learning on the model's performance.

Conversely, removing the instructional component had no noticeable effect on performance. The results were identical to those from the second main experiment, with the model correctly classifying the same three fraudulent cases and five non-fraudulent cases. These findings indicate that while few-shot learning proves to be an effective prompt engineering method, instructional learning did not enhance performance. This ineffectiveness can be attributed to several factors. ChatGPT does not have access to external resources that were not used during its initial training. References to specific papers, metrics, or resources like the Loughran and McDonald dictionary are unhelpful if these are unknown and inaccessible to the model. Moreover, general instructions about the importance of verb quantity, or the use of emotion words, pronouns, and exclusive terms did not yield positive results. Perhaps, direct commands such as “count and compare the average number of verbs in fraudulent and legitimate texts and use it as a predictor” might be more effective, though this approach would reduce the model to a basic NLP tool.

6. Conclusions

The falsification of financial statements is a well-documented form of fraud and represents the most costly type of occupational fraud. Given its severe repercussions, the detection of falsified financial statements has garnered significant attention over the past three decades. Previous studies have utilized statistical, machine learning, and more recently, deep learning methods to develop models capable of identifying fraudulent cases. These studies have analyzed both quantitative-numerical and qualitative-textual data. Regarding qualitative data, research findings indicate that linguistic analysis of text segments extracted from annual reports can successfully identify falsified financial statements. In the past three years, Large Language Models (LLMs) have rapidly gained immense popularity. LLMs are well-versed in linguistic structures and characteristics, and have been effectively applied to tasks such as sentiment analysis, translation, summarization, and more. Beyond these general capabilities, LLMs can acquire specialized knowledge and skills through fine-tuning and, most notably, prompt engineering. This involves providing natural language instructions as inputs that guide the model in performing specific tasks. Their advanced reasoning abilities, knowledge of language patterns, and natural language communication skills make LLMs invaluable tools for accountants and auditors. Researchers highlight the potential of LLMs in audit-related tasks, including financial ratio analysis, journal entry testing, risk assessment, and the automation of audit processes.

In this study, we evaluate the capability of a Large Language Model (LLM), specifically ChatGPT-4, to detect falsified financial statements through linguistic analysis of text segments extracted from annual reports. We focus our analysis on the CEO's letter to stakeholders and the sections discussing the risks the company faces. Our approach treats the LLM as an end-to-end fraud detection tool rather than merely a feature extractor. Additionally, we choose to rely exclusively on prompt engineering, avoiding fine-tuning—a practice that typically requires the expertise of machine learning specialists. The central question of our study is whether auditors, who are experts in their field but not in machine learning, can effectively utilize an LLM by crafting natural language prompts to aid in the identification of fraud within annual reports.

We conducted several experiments to assess the capabilities of ChatGPT in detecting fraudulent financial statements. The results revealed that the model's inherent knowledge and capabilities were insufficient for identifying fraud cases, as it initially classified all texts as legitimate. However, the application of prompt engineering techniques enabled the model to distinguish between fraudulent and legitimate cases to some extent. Notably, few-shot learning had a significantly beneficial impact, whereas instructional learning did not meaningfully influence the model's performance. By incorporating human feedback into the model's classification process, we enhanced its performance in terms of Sensitivity and F-Measure. This improvement indicates that the model became more effective in detecting cases of falsification. These findings highlight the potential of Large Language Models (LLMs) in identifying financial statement fraud. Regarding the data used, the CEO's letter to shareholders proved to be informative and effective for differentiating between fraudulent and legitimate cases. Conversely, the text segment related to risks was found to be unsuitable for this task. Enriching the input with risk-relevant text significantly worsened the model's performance, suggesting that incorporating non-informative text confuses the model.

We acknowledge some limitations in the current study. Previous research on financial statement fraud detection has reported better performance scores, but these studies typically employed conventional machine learning techniques. Such approaches require a deep understanding of machine learning methodologies and the development of a sufficiently large training/testing dataset, which is both time-consuming and labor-intensive. Notably, Large Language Models (LLMs) are expanding rapidly in size and scope. This expansion suggests that future models may possess a broader knowledge base and could be trained with additional data pertinent to financial statement fraud detection. Access to specific resources, such as the Loughran and McDonald dictionary or advanced NLP techniques, could be facilitated through LLMs. Additionally, ongoing research in prompt engineering is proving to be intense and productive, with new, sophisticated prompting methods likely to enhance performance further. For now, we highlight the potential of LLMs to detect fraudulent cases and posit that significant performance improvements can be expected in the near future. Another consideration is the relatively small sample size used in our study. The principle of few-shot learning suggests that only a limited number of examples are necessary to train the model effectively. We employed six examples for training, which can be deemed adequate. However, incorporating more cases in the testing phase could potentially refine the assessment.

This study can serve as a foundational step for future research. One potential avenue involves instruction learning with more precise and direct instructions, compelling the model to calculate specific metrics. The Chain-of-Thoughts (CoT) is a promising prompting technique that structures examples as input–chain of thoughts–output. To date, CoT has predominantly been applied to relatively straightforward problems, particularly in arithmetic reasoning. For financial statement fraud detection, this technique could be adapted to include the CEO's letter, a detailed reasoning path for text analysis, and the desired output, testing its effectiveness in this specific domain. Additional prompting techniques, such as Tree-of-Thoughts, could also be explored. Moreover, designing specific prompts tailored for the task of fraudulent financial statement identification could enhance detection capabilities. These topics present intriguing challenges for subsequent research endeavors.

References

Achakzai, M. a. K., & Peng, J. (2023). Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis (Online)/International Review of Financial Analysis*, *89*, 102827. https://doi.org/10.1016/j.irfa.2023.102827

(2020). *ACFE 2020 Report to the Nations*. https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf

Ashtiani, M. N., & Raahemi, B. (2022). Intelligent Fraud Detection in Financial Statements Using Machine learning and Data Mining: A Systematic Literature review. *IEEE Access*, *10*, 72504–72525. https://doi.org/10.1109/access.2021.3096799

Beasley, M .S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting Review 71*(4), 443-465.

Beasley, M. S., Carcello, J. V., Hermanson, D. R., & Lapides, P. D. (2000). Fraudulent Financial Reporting: consideration of industry traits and corporate governance mechanisms. *Accounting Horizons*, *14*(4), 441–454. https://doi.org/10.2308/acch.2000.14.4.44

Bell, T. B., & Carcello, J. V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing*, *19*(1), 169–184. https://doi.org/10.2308/aud.2000.19.1.169

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Bhattacharya, I., & Mickovic, A. (2024). Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems*, *53*, 100682. https://doi.org/10.1016/j.accinf.2024.100682

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164–175. https://doi.org/10.1016/j.dss.2010.07.012

Chen, Y., Yang, G., Wang, D., & Li, D. (2024). Eliciting knowledge from language models with automatically generated continuous prompts. *Expert Systems With Applications*, *239*, 122327. https://doi.org/10.1016/j.eswa.2023.122327

Cheng, C., Kao, Y., & Lin, H. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Applied Soft Computing*, *108*, 107487. https://doi.org/10.1016/j.asoc.2021.107487

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements*. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21–41. https://doi.org/10.1002/(sici)1099-1174(199803)7:1

Fotoh, L., & Mugwira, T. (2023). The use of ChatGPT in External Audits: Implications and future research. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4453835

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.29

Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, *50*(3), 595–601. https://doi.org/10.1016/j.dss.2010.08.010

Goel, S., Gangolly, J., Faerman, S. R., & Uzuner, O. (2010). Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting*, *7*(1), 25–46. https://doi.org/10.2308/jeta.2010.7.1.25

Goel, S., & Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *23*(3), 215–239. https://doi.org/10.1002/isaf.1392

Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing*, *16*(1), 14. https://www.questia.com/library/journal/1G1-20745791/assessing-the-risk-of-management-fraud-through-neural

Gu, H., Schreyer, M., Moffitt, K., & Vasarhelyi, M. A. (2023). Artificial Intelligence Co-Piloted Auditing. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4444763

Guan, H., Li, S., Wang, Q., Lyulyov, O., & Pimonenko, T. (2022). Financial fraud identification of the companies based on the logistic regression model. *Journal of Competitiveness*, *14*(4). https://doi.org/10.7441/joc.2022.04.09

Gunawardena, C. N., Chen, Y., Flor, N., & Sanchez, D. (2023). Deep learning models for analyzing social construction of knowledge online. *Online Learning*, *27*(4). https://doi.org/10.24059/olj.v27i4.405

Gupta, S., & Mehta, S. K. (2021). Data Mining-based Financial Statement Fraud Detection: Systematic literature review and meta-analysis to estimate data sample mapping of fraudulent companies against non-fraudulent companies. *Global Business Review*, 097215092098485. https://doi.org/10.1177/0972150920984857

Hillebrand, L., Berger, A., Deußer, T., Dilmaghani, T., Khaled, M., Kliem, B., Loitz, R., Pielka, M., Leonhard, D., Bauckhage, C., & Sifa, R. (2023). Improving Zero-Shot Text Matching for Financial Auditing with Large Language Models. In *The 23rd ACM Symposium on Document Engineering*. https://doi.org/10.1145/3573128.360934

Hu, K. (2023). *ChatGPT sets record for fastest-growing user base—analyst note*. Reuters. https://www.reuters.com/technology/chatgpt-setsrecord-fastest-growing-user-base-analyst-note-2023-02-01/

Huang, L., Abrahams, A., & Ractham, P. (2022). Enhanced financial fraud detection using cost-sensitive cascade forest with missing value imputation. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *29*(3), 133–155. https://doi.org/10.1002/isaf.1517

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009

Karpathy, A. @karpathy. (2023). *The hottest new programming language is English*. Twitter. https://twitter.com/karpathy/status/1617979122625712128?s=20

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems With Applications*, *32*(4), 995–1003. https://doi.org/10.1016/j.eswa.2006.02.016

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.11916

Lee, T. A., Ingram, R. W., & Howard, T. P. (1999). The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud*. *Contemporary Accounting Research*, *16*(4), 749–786. https://doi.org/10.1111/j.1911-3846.1999.tb00603.x

Li, H., De Freitas, M. M., Lee, H., & Vasarhelyi, M. (2024). Enhancing Continuous Auditing with Large Language Models: A Framework for Cross-Verification Using Exogenous Textual Data. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4692960

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A systematic survey of prompting methods in natural language processing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2107.13586

Long, J. (2023). Large language model guided Tree-of-Thought. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.08291

Minhas, S., & Hussain, A. (2016). From spin to swindle: Identifying falsification in financial text. *Cognitive Computation*, *8*(4), 729–745. https://doi.org/10.1007/s12559-016-9413-9

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023, November 28). *Can Generalist Foundation models outcompete Special-Purpose Tuning? Case study in medicine*. arXiv.org. https://arxiv.org/abs/2311.16452

Omar, N., Johari, Z.A., & Smith, M. (2017). Predicting fraudulent financial reporting using artificial neural network. *Journal of Financial Crime*, *24*(2), 362–387. https://doi.org/10.1108/jfc-11-2015-0061

Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., & Ribeiro, M. T. (2023). ART: Automatic multi-step reasoning and tool-use for large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.09014

Perols, J. (2011). Financial Statement Fraud Detection: An analysis of statistical and machine learning algorithms. *Auditing*, *30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Purda, L., & Skillicorn, D. (2014). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, *32*(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089

Ramlochan, S. (2023, May 27). *Role-Playing in Large Language Models like ChatGPT*. Prompt Engineering. https://promptengineering.org/role-playing-in-large-language-models-like-chatgpt/

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500. https://doi.org/10.1016/j.dss.2010.11.006

Riskiyadi, M. (2023). Detecting future financial statement fraud using a machine learning model in Indonesia: a comparative study. *Asian Review of Accounting*. https://doi.org/10.1108/ara-02-2023-0062

Shahana, T., Lavanya, V., & Bhat, A. R. (2023). State of the art in financial statement fraud detection: A systematic review. *Technological Forecasting & Social Change/Technological Forecasting and Social Change*, *192*, 122527. https://doi.org/10.1016/j.techfore.2023.122527

Shin, T., Razeghi, Y., Logan, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.15980

Soltani, M., Kythreotis, A., & Roshanpoor, A. (2023). Two decades of financial statement fraud detection literature review; combination of bibliometric analysis and topic modeling approach. *Journal of Financial Crime*, *30*(5), 1367–1388. https://doi.org/10.1108/jfc-09-2022-0227

Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, *17*(4), 179–191. https://doi.org/10.1108/02686900210424321

Summers, S.L., & Sweeney, J.T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis. *Accounting Review, 73*(1), 131-146.

Vasarhelyi, M. A., Moffitt, K. C., Stewart, T., & Sunderland, D. (2023). Large Language Models: an emerging technology in accounting. *Journal of Emerging Technologies in Accounting*, *20*(2), 1–10. https://doi.org/10.2308/jeta-2023-047

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency improves chain of thought reasoning in language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.11171

Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, *167*, 113913. https://doi.org/10.1016/j.dss.2022.113913

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought prompting elicits reasoning in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2201.11903

Willey, L., White, B.J., and Deale, C.S. (2023). Teaching AI in the college course: introducing the AI prompt development life cycle. *Issues in Information Systems, 24*(2), pp. 123-138. DOI: https://doi.org/10.48009/2_iis_2023_111

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.10601

Yadav, A. K. S., & Sora, M. (2022). Unsupervised learning for financial statement fraud detection using manta ray foraging based convolutional neural network. *Concurrency and Computation*, *34*(27). https://doi.org/10.1002/cpe.7340

Zhang, Y., Hu, A., Wang, J., & Zhang, Y. (2022). Detection of fraud statement based on word vector: Evidence from financial companies in China. *Finance Research Letters*, *46*, 102477. https://doi.org/10.1016/j.frl.2021.102477

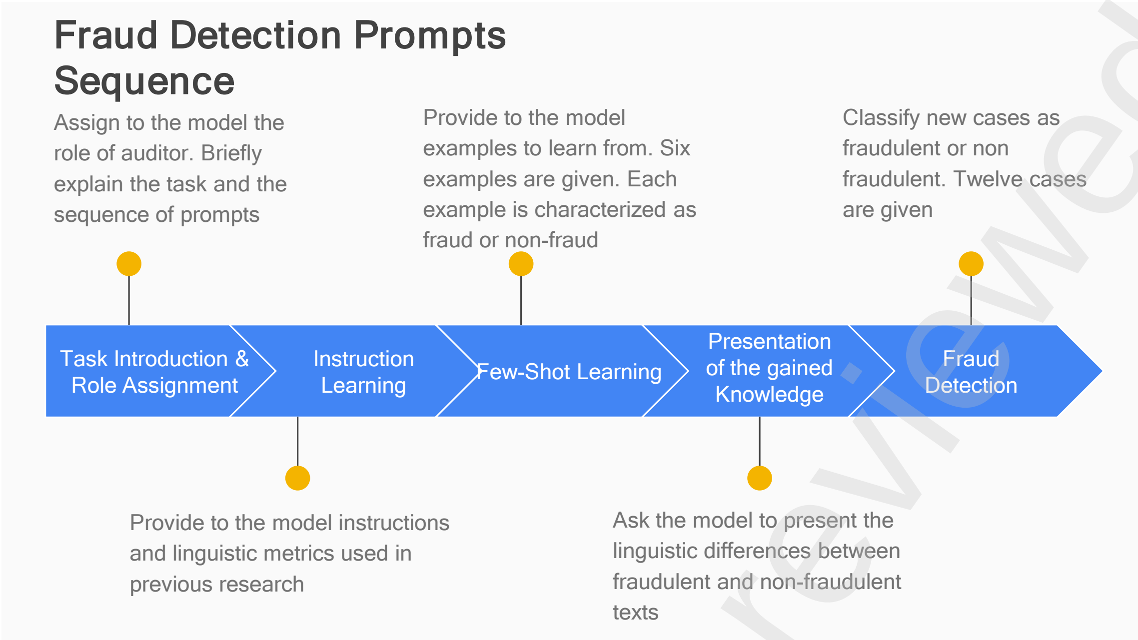

Figure 1. The sequence of prompts

Appendix A.

Firms included in the experiment

Firms involved in financial statement fraud. The first three firms have been used for Few-Shot Learning. The rest firms have been used for testing.

Count	Firm	Year
1	Huishan	2016
2	Olympus	2010
3	Noble Group	2016
4	Sino – Forest Corporation	2009
5	Steinhoff	2016
6	Tesco	2013
7	Toshiba	2014
8	Valeant Pharmaceuticals	2014
9	Wirecard	2017

Matching firms not involved in financial statement fraud. The first three firms have been used for Few-Shot Learning. The rest firms have been used for testing.

Count	Firm	Year	Matching
1	China Mengniu	2016	Huishan
2	Nikon	2010	Olympus
3	Archer Daniels Midland	2016	Noble Group
4	Weyerhaeuser	2009	Sino – Forest Corporation
5	Wayfair	2016	Steinhoff
6	Carrefour	2013	Tesco

7	Sony	2014	Toshiba
8	Perrigo	2014	Valeant Pharmaceuticals
9	Paypal	2017	Wirecard