Leiden University

ICT in Business and the Public Sector

Evaluating LLMs in Practice: Enhancing Going Concern Assessments for Healthcare Organizations

Name:	Tessa de Jong
Student ID:	s2553082

Date:	14/04/2025
1st supervisor:	Marc Hilbert
2nd supervisor:	Joost Visser
Company supervisor:	Marco Prinsen (BDO)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: The COVID-19 pandemic intensified already existing financial uncertainty, especially within healthcare organizations, highlighting the need for effective going concern assessments to predict financial distress. Traditional machine learning methods have proven effectiveness in predicting bankruptcy, but face challenges in analyzing complex and unstructured data. Large language models (LLMs) have shown promise in processing these complex data types, yet their application in auditing, particularly for assessing the going concern of healthcare organizations, remains unexplored.

Aim: This research aims to show the potential of LLMs to predict bankruptcy of healthcare organizations. By integrating structured and unstructured information, the study seeks to evaluate the effectiveness of LLMs in improving the going concern assessment in existing auditing processes. The focus lies on assessing the model's performance in practice.

Method: We followed a design science approach to develop a LLM-based model for assessing going concern in healthcare organizations, using the DSPy framework to programmatically optimize the LLM. The model integrated structured financial data and unstructured audit reports to improve the accuracy of bankruptcy prediction and reduce assessment time. Its effectiveness was evaluated through a case study comparing prediction accuracy and time efficiency against traditional assessment methods used by accountants. In this study, the control group received only financial data and a single page of the auditor's report, while the LLM-assisted group also received the model's prediction and a model-generated summary of the same auditor's report page. Additionally, we collected qualitative feedback on the usability of and satisfaction with the model.

Results: The results showed that the model achieved an accuracy of up to 79%, with DSPy's COPRO optimizer for prompt optimization emerging as the best-performing option due to its high accuracy (79%) and low false negative rate (10). Feature selection significantly impacts model accuracy, with expert-selected features outperforming mutual information-based selection, suggesting the importance of domain expertise in bankruptcy prediction. In the case study, the model matched or outperformed individual accountants in predicting bankruptcy. The LLM-assisted group demonstrated a slight improvement in assessing edge cases compared to the control group. Despite no measurable difference in overall accuracy or assessment speed, accountants found the model useful for highlighting key financial indicators, with all participants stating they would incorporate it into their real-world evaluations.

Conclusion: This research demonstrated the potential of LLMs in enhancing the going concern assessments for healthcare organizations. While LLMs did not outperform traditional methods in terms of accuracy or efficiency, their ability to support auditors by identifying key financial indicators and providing early warnings proves valuable. The accountants showed a positive attitude towards the model, highlighting its usefulness as an assistive tool in their workflow by complementing existing auditing processes rather than replacing them. Future work should focus on refining the model, expanding its use across sectors, and effectively integrating the model into audit processes.

Acknowledgements

First of all, I would like to thank my first supervisor, Marc Hilbert, for his support throughout this entire process. He believed in this project from the start and his enthusiasm was inspiring. While every student experiences some challenges during their thesis, these were kept minimal due to the feedback and help Marc gave along the way. I truly believe that he had a big part in making this experience not only manageable, but even enjoyable at times.

Secondly, I would like to express my gratitude to my company supervisor, Marco Prinsen. Despite his already full schedule, he always made time to discuss my thesis and offer help in any way he could. It reflects the support of the company itself, as not only him, but also many other colleagues offered their advice and feedback. This welcoming environment made a significantly positive difference in my experience.

Additionally, I would like to thank Joost Visser, my second supervisor. He showed genuine interest in the project, motivating me to get the most out of my thesis. His questions were always thoughtful and critical, which challenged me to refine my work and made the research into what it is today.

Lastly, I am grateful for the support of my family and friends. They were always happy to hear me voice my worries, which were, of course, never needed, and kept me with both feet on the ground.

Contents

Abstract . . . 1
Acknowledgements . . . 1

1 Introduction **3**
1.1 Problem statement . . . 3
1.2 Research question . . . 3
1.3 Business case . . . 4
1.4 Overview of the thesis . . . 4

2 Background and related work **5**
2.1 Going concern assumption . . . 5
2.2 Auditing: traditional approaches to going concern assessment . . . 5
2.3 Healthcare organizations & going concern . . . 6
2.4 Emerging role of AI for going concern . . . 7
2.5 The use of LLMs in auditing . . . 9

3 Method **11**
3.1 Data collection and pre-processing . . . 12
3.2 Feature selection . . . 13
3.3 Sampling . . . 13
3.4 Model optimization with DSPy . . . 14
3.4.1 Aspects of the DSPy framework . . . 14
3.4.2 Optimizers . . . 15
3.5 Boundaries of the model . . . 18
3.6 Case Study . . . 19

4 Results **21**
4.1 Model accuracy . . . 21
4.2 Boundaries of the model . . . 26
4.3 Case study . . . 27
4.3.1 Accuracy . . . 27
4.3.2 Speed . . . 27
4.3.3 Ease of assessment . . . 29
4.3.4 Satisfaction with the model . . . 29

5 Discussion **31**
5.1 Model accuracy . . . 31
5.1.1 Feature selection . . . 31
5.1.2 Optimizers . . . 31
5.1.3 COPRO optimizer . . . 32
5.2 Boundaries of the model . . . 33
5.3 Case study . . . 33
5.3.1 Difference between model and accountants . . . 33
5.3.2 Difference between control and LLM group . . . 36
5.3.3 Satisfaction with model . . . 38
5.4 Hypotheses . . . 38
5.5 DSPy framework . . . 39

5.6 Limitations . . . 40
5.7 Recommendations . . . 41
5.7.1 Model design recommendations . . . 41
5.7.2 Model integration recommendations . . . 41

6 Conclusions **43**
6.1 Summary . . . 43
6.2 Answering the research question . . . 44
6.3 Contributions . . . 44
6.4 Future work . . . 45

A Feature selection **50**

B Full table of model results **51**

C Case study **53**
C.1 Case study form . . . 53
C.2 Evaluation form . . . 57

Chapter 1

Introduction

1.1 Problem statement

The COVID-19 pandemic significantly impacted healthcare organizations, including hospitals, nursing homes, and other care facilities, and caused substantial financial losses and operational disruptions. Healthcare centers faced revenue declines, workforce reductions, and shortages in medical supplies, which intensified the financial pressures they were already under [56]. Along with a volatile global situation and increased competition, the need for accurate assessment of going concern risk became even more critical [71].

Going concern assessments play a key role in determining the financial health of organizations, especially within the field of auditing. In this context, audits are "independent examinations of the records of an organization to ascertain how far the financial statements present a true picture of the firm and that accounting systems are well-controlled, legal, and accurate" [70]. Accurate and efficient auditing processes are essential for identifying potential risks, such as default or operational disruptions, and guiding organizations in mitigating these going concern risks [60].

While machine learning has been widely applied to enhance the going concern assessments, these methods often struggle to process complex, high-volume data [43]. Large language models (LLMs), however, offer the potential to address this gap. Unlike traditional machine learning models, LLMs are designed to process and analyze both structured and unstructured data, allowing them to uncover patterns in text that might be missed by conventional approaches [9]. However, their application in assessing going concern risk needs to be evaluated and compared with traditional approaches to assess their practical value [37]. Moreover, LLMs have yet to be fully explored in the context of healthcare organizations. Healthcare-related audit research remains sparse in general [72], but especially when it comes to integrating advanced technologies like LLMs into traditional auditing processes.

Therefore, this research aims to empirically evaluate the effectiveness of LLMs in assessing going concern risks by analyzing both structured and unstructured data from Dutch healthcare organizations. By comparing the performance of LLMs to traditional audit methods, this study seeks to uncover the strengths and limitations of LLMs in practice. Ultimately, this study will contribute to the growing literature on LLMs in auditing by providing practical guidance on how these models can be applied to going concern assessments of the healthcare sector.

1.2 Research question

Following the problem statement, the research question is stated as:

"How can LLMs be utilized and optimized effectively to enhance the going concern assessment for healthcare organizations?"

Addressing this research question, the research will be divided into the following hypotheses:

H.1 Programmatic prompt-tuned LLMs are more effective than traditional auditing methods in assessing the going concern.

Effectiveness will be measured by comparing prediction accuracy of the assessment.

H.2 Programmatic prompt-tuned LLMs can shorten the current process of assessing the going concern while maintaining accuracy.
This will be evaluated by measuring the time taken by accountants for the going concern assessment with and without the model.

H.3 Accountants perceive LLMs as a valuable tool to assist in going concern assessments.
An evaluation form will be used to assess accountants' perceptions of the model's usefulness and ease of integration into their workflow.

1.3 Business case

This research was carried out in collaboration with BDO Netherlands. BDO is a network of advisory firms that provide a wide range of services including accountancy, business advice, tax, and digital transformation [8]. It is the fifth-largest accounting network in the world [13].

In the context of an accounting firm like BDO, going concern assessments are crucial. These assessments are an integral part of the audit and involve evaluating a company's ability to continue operating in the foreseeable future. For BDO's audit services, ensuring the going concern of clients is essential to meet the expectations of key stakeholders and maintain the integrity of financial reporting.

This study aims to use a LLM to enhance the efficiency and accuracy of the going concern assessments, empowering accountants with more precise and data-driven recommendations to clients. This aligns with BDO's commitment to providing smart solutions and new insights to their clients [7].

BDO supplied the dataset mentioned in Chapter 3 and supported this research by offering access to a network of accountants. These professionals provided insights into going concern assessment processes and contributed to the case study described later.

1.4 Overview of the thesis

The remainder of this thesis will be structured as follows. We start by establishing the theoretical background in Chapter 2, where we provide the literature on going concern assessments and the related work regarding the use of LLMs in this context. Chapter 3 outlines the methodology used to answer the research question, with the results presented in Chapter 4. In Chapter 5, we interpret the findings, relate them to the hypotheses, discuss the limitations of the study, and offer recommendations for the host company. Lastly, Chapter 6 concludes the thesis with a summary, an answer to the research question, contributions to the field, and suggestions for future work.

Chapter 2

Background and related work

The following chapter presents the academic literature on going concern assessments and the use of LLMs in auditing. It begins by defining the going concern assumption and its assessment process. Subsequently, we examine the distinctions between healthcare organizations and traditional for-profit entities, as well as the impact on the going concern assessment. Finally, the emergence of AI and LLM applications in auditing is outlined. This chapter is the foundation for our study on the use of LLMs in going concern assessments.

2.1 Going concern assumption

What is the going concern assumption?

The going concern assumption is a fundamental concept in both international and national accounting standards. The International Accounting Standards Board (IASB), which applies in regions that have adopted International Financial Reporting Standards (IFRS) — such as the European Union, Australia, and many parts of Asia and Africa — includes this assumption in IAS 1. IAS 1 states that organizations are expected to continue their operations as a going concern [14]. Should an auditor, during their assessment, identify uncertainties that cause significant doubt of the organization's ability to continue as a going concern, these uncertainties must be disclosed [14].

Similarly, under U.S. Generally Accepted Accounting Principles (GAAP), the going concern assumption is a standard for financial statement presentation unless information suggests otherwise. Since 1989, auditors in the U.S. have been required under Statement on Auditing Standards (SAS) No. 59 to assess whether substantial doubt exists regarding the ability of an organization to continue as a going concern for the next twelve months [24]. If there is any doubt, auditors must modify their audit reports accordingly to reflect the uncertainty [59].

Both IASB and U.S. GAAP stress the importance of assessing and disclosing going concern risks, ensuring that financial statements accurately reflect an organization's ability to operate.

2.2 Auditing: traditional approaches to going concern assessment

How is going concern currently assessed within auditing?

A going concern assessment involves the auditor's evaluation of any uncertainties that could threaten the organization's ability to continue operations, in line with the aforementioned going concern assumption [15]. They assess the risk that a company will need to cease their operations within the next year.

As Cormier et al. [24] explain, the going concern concept involves more than just the risk of bankruptcy. While this is often the main focus, other significant events, such as restructuring, (partial) liquidation, or a merger, may also indicate that a company is no longer able to continue its operations in the same manner [24]. This broader interpretation ensures that the auditor evaluates various financial difficulties that could influence the company's continuity, beyond the sole risk of default. For example, if a company is undergoing significant restructuring, this may indicate potential liquidity issues that could affect its long-term continuity [6]. However, since bankruptcy provides definitive prove that an organization is no longer a going concern, this study will focus on it as the primary target for assessing going concern risks. By predicting bankruptcy, we can address the most concrete and impactful outcome within continuity assessments. Therefore, for the purpose of this thesis, the going concern assessment will be considered equivalent to bankruptcy prediction.

Mutchler [52] researched in 1984 what auditors look for when assessing going concern in practice. They found that the top five ratios were: cash flow from operations/total debt, current assets/current liabilities, net worth/total debt, total debt/total assets, and total liabilities/total assets. A later study from 1996 found similar ratios, with the key differences being the inclusion of change in net worth/total liabilities and the replacement of net worth/total debt with net worth/total liabilities [41]. These ratios are particularly relevant as they reveal aspects of liquidity, solvency, and the organization's ability to manage factors that directly indicate financial distress, impacting the overall going concern.

However, as Carson et al. [19] said: "As the audit environment changes, there is an ongoing need to update evidence on what financial statement variables auditors rely on in practice when making [going concern] decisions." Especially in a time where there is more and more data available, this highlights the importance of continuous research to identify and validate the key factors necessary for accurate going concern assessments. But before exploring these factors further, it is essential to clarify how going concern assessments differ between healthcare organizations and typical profit organizations.

2.3 Healthcare organizations & going concern

Why is going concern in healthcare organizations important and how is it different from other organizations?

Healthcare organizations include a wide range of entities, such as hospitals, nursing homes, rehabilitation centers, and outpatient care providers, that play a vital role in delivering essential health services to the public. A failed going concern assessment in this sector could disrupt an organization's critical services, affecting public welfare and access to healthcare. The uncertainty brought by external crises like the pandemic further underscores the need for reliable assessments to maintain financial stability and public health [56].

In the Netherlands, healthcare organizations are required by law to publish their annual report (Wet marktordening gezondheidszorg, article 40b) [45]. The law ensures accountability in healthcare, allowing stakeholders to assess the continuity and professionalism of an organization, as well as their potential risks [46]. This requirement underlines the importance of transparency in financial reporting. The publication of annual reports allows auditors and stakeholders to effectively evaluate these organizations' financial health and compare them within the same context.

However, healthcare organizations tend to require a distinct approach when it comes to going concern assessments. For example, in the Netherlands, medical services aimed at personal healthcare for individuals are exempt from VAT [10]. Moreover, accounting in healthcare research is "directly associated with governmental objectives to preserve human rights" [72]. These objectives have an impact on the different regulations and potential subsidies healthcare organizations receive. Furthermore, many dutch healthcare organizations operate as non-profits [69], affecting how taxes and revenues are managed compared to for-profit entities. It also indicates a trade-off between quality of care and financial metrics like cost reduction and efficiency [58]. For instance, non-profit status can lead to different funding opportunities, which may not prioritize immediate profitability but rather long-term sustainability and service quality [30]. In addition, healthcare revenue streams are complex due to the interaction of various payment models and the need for collaboration among multiple stakeholders in the healthcare system [67].

These characteristics create a unique financial and operational landscape in the healthcare sector, that need to be kept in mind while assessing their going concern.

2.4 Emerging role of AI for going concern

How has AI been used to enhance going concern assessments?

As mentioned before, it is crucial to continuously research the factors needed for accurate going concern assessments, as the audit environment continues to change and evolve [19]. Multiple studies have researched the factors included in going concern assessments and bankruptcy predictions. One of the most prominent studies has been done by Altman [2], who formulated the Altman Z-score. This score takes financial ratios to predict financial distress in the following manner:

$$Z = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E \tag{2.1}$$

Where:

$$A = \frac{\text{Working capital}}{\text{Total assets}}$$
$$B = \frac{\text{Retained earnings}}{\text{Total assets}}$$
$$C = \frac{\text{Earnings before interest and taxes}}{\text{Total assets}}$$
$$D = \frac{\text{Market value of equity}}{\text{Book value of total debt}}$$
$$E = \frac{\text{Sales}}{\text{Total assets}}$$

The Altman Z-score is widely used to assess the financial health of a company and predict its likelihood of bankruptcy, with applications in various industries [3]. In a later study, Altman and McGough [4] found that their failure prediction model even outperformed the auditors when it came to the going concern assessment of a set of American organizations.

Building on Altman's work, Bellovary et al. [12] conducted a literature research on 165 bankruptcy prediction studies from 1930 to 2004. They focused on the factors used within these studies and listed those that were mentioned in five or more studies (see Figure 2.1).

While these factors are widely used across industries, healthcare-specific characteristics can significantly influence financial distress. Landry and Landry III [40] conducted a study examining the factors associated with hospital bankruptcies in the US. Their findings suggest that financially distressed hospitals tend to be smaller in size, less likely to be affiliated with a hospital network, and more often investor-owned. Building on this, Liu et al. [42] identified additional factors influencing hospital financial distress and closure. Their study found that for-profit ownership, high debt ratios, a larger proportion of Medicare and Medicaid patients, urban location, and teaching hospital status increase financial risk. Conversely, hospitals with higher occupancy rates, stronger financial performance (e.g., return on investment and asset turnover), and affiliations with multihospital systems were less likely to file for bankruptcy. Additionally, hospitals designated as Sole Community, Medicare Dependent, or Small Rural had a lower likelihood of failure [42]. These factors highlight the unique risks hospitals face and provide insight into financial distress within the healthcare sector. However, more general bankruptcy prediction models consider a wider range of variables, as shown in Figure 2.1. Although healthcare-specific factors are essential for understanding financial distress in this sector, integrating both general and sector-specific variables allows for a more comprehensive analysis.

The way these factors are used in going concern assessments varies between the individual studies. Up until the 1990s, the most used models to evaluate these variables were discriminant analyses [12], as the Altman Z-score developed by Altman [2]. However, with recent advances in technology, these traditional financial models are now enhanced by machine learning methods that can identify complex patterns in financial data. This shift has introduced a new depth of analysis in financial risk prediction.

Appendix B
Factors Included in Five or More Studies[5]

Factor/Consideration	Number of Studies that Include
Net income / Total assets	54
Current ratio	51
Working capital / Total assets	45
Retained earnings / Total assets	42
Earnings before interest and taxes / Total assets	35
Sales / Total assets	32
Quick ratio	30
Total debt / Total assets	27
Current assets / Total assets	26
Net income / Net worth	23
Total liabilities / Total assets	19
Cash / Total assets	18
Market value of equity / Book value of total debt	16
Cash flow from operations / Total assets	15
Cash flow from operations / Total liabilities	14
Current liabilities / Total assets	13
Cash flow from operations / Total debt	12
Quick assets / Total assets	11
Current assets / Sales	10
Earnings before interest and taxes / Interest	10
Inventory / Sales	10
Operating income / Total assets	10
Cash flow from operations / Sales	9
Net income / Sales	9
Long-term debt / Total assets	8
Net worth / Total assets	8
Total debt / Net worth	8
Total liabilities / Net worth	8
Cash / Current liabilities	7
Cash flow from operations / Current liabilities	7
Working capital / Sales	7
Capital / Assets	6
Net sales / Total assets	6
Net worth / Total liabilities	6
No-credit interval	6
Total assets (log)	6
Cash flow (using net income) / Debt	5
Cash flow from operations	5
Operating expenses / Operating income	5
Quick assets / Sales	5
Sales / Inventory	5
Working capital / Net worth	5

Figure 2.1: Factors included in five or more bankruptcy prediction studies from 1930-2004 [12].

One of the earliest studies in this area was conducted by Bell et al. [11], who compared logistic regression with a neural network, finding that both models performed equally well in predicting bank failures. Nevertheless, three years later, another study demonstrated that their neural network had a higher accuracy in predicting financial distress than a Multiple Discriminant Analysis (MDA) [23]. The power of neural networks was again shown by Ciampi and Gordini [21] for small enterprises and by Wu et al. [71] for listed Chinese companies. As Krulicky and Horak [39] said, the uncertain environment, new forms of business, and the need for handling substantial amounts of data, along with sufficient computing resources and reliable software tools, make artificial neural networks some of the most effective models for evaluating and predicting business development.

Not only neural networks have shown their capabilities within the literature. Yuan et al. [73] proposed a two-stage model combining k-means clustering with support vector domain description to predict company defaults up to five years in advance. Their research demonstrated the superiority of multi-stage machine learning models, particularly in identifying financial indicators critical for predicting bankruptcy. Similarly, Sigrist and Leuenberger [66] introduced the "LaGaBoost frailty model", integrating tree-boosting with a latent frailty model to predict default probabilities, showcasing the importance of capturing unobservable correlations in risk prediction. Moradi et al. [47] took another approach and used Fuzzy Clustering Means to predict going concern of Iranian companies, which obtained good results for financial distress prediction. In a different study, Moscatelli et al. [48] found that machine learning models significantly enhance default risk prediction using limited financial data compared to traditional statistical models, although this advantage diminishes when high-quality information and small datasets are involved. They demonstrated this difference in accuracy by using random forest and gradient-boosted tree models, while using traditional statistical methods as benchmarks [48].

All in all, recent studies emphasize the effectiveness of machine learning methods in enhancing default risk prediction, particularly when relying on limited financial data. These advancements showcase the ability of machine learning models to uncover patterns and relationships within financial data, reinforcing their value in going concern assessments.

2.5 The use of LLMs in auditing

What are Large Language Models and how can they be used for going concern assessments?

Even though the aforementioned studies have shown promising results, these traditional machine learning methods struggle to effectively process complex and high-volume data [43]. In addition, these studies focus primarily on structured financial data. That is why recent research has begun to explore the integration of unstructured data sources, which is challenging for machine learning methods [43].

The value of focusing on the integration of unstructured data sources can be deduced from research conducted by Muñoz-Izquierdo et al. [51]. They found that including the audit report in an accounting model results in a significantly higher accuracy than using solely the accounting model. This indicates that using the unstructured data from the audit report can add value to the prediction model. Another study by the same author suggests that audit reports can also provide insights into the causes behind an organization's bankruptcy [50]. The ability to include unstructured data in models is particularly beneficial for healthcare organizations, where data can be complex and multi-layered, as described in Section 2.3.

In contrast to machine learning methods, large language models are capable of processing and summarizing unstructured data [9]. Language Models are defined as "statistical models which assign a probability to a sequence of words" [18], where Large Language Models (LLMs) are the larger version of these models with significantly increasing amounts of model parameters and larger training datasets [53]. LLMs are based on the transformer architecture, enabling them to excel in tasks like text generation, machine translation, question answering, and sentiment analysis [53]. Transformers use self-attention, which allows them to analyze the entire input sequence at once, rather than processing it incrementally [29]. Using this mechanism, each word or token assesses the relevance of every other word in the sequence, effectively capturing long-range dependencies and contextual relationships [29]. This flexibility is especially helpful when dealing with varied and complex data, like the financial state of healthcare organizations.

Another key advantage of LLMs over traditional machine learning models is their ability to perform relatively well on new tasks without explicit training on specific datasets [16]. While conventional machine learning models are trained in supervised settings for specific tasks, LLMs are developed in a self-supervised manner on large amounts of documents, in order for them to learn generalized representations applicable to various tasks [53]. This capability allows LLMs to perform new tasks without the need for fine-tuning on large amounts of training data. This is particularly valuable in areas with limited data, such as bankruptcy prediction.

LLMs have demonstrated significant potential in financial analysis, particularly for extracting insights from unstructured data. A promising model in the financial domain is FinBERT, a variant of the BERT model trained specifically on financial texts [33]. FinBERT has produced higher accuracy in tasks like sentiment analysis by understanding the complexities of financial language. Its ability to extract meaningful insights from financial reports can be beneficial for organizations by enabling more accurate decision-making [33].

In addition to sentiment analysis, LLMs have also been evaluated for financial statement analysis. For instance, Cao et al. [17] developed the RiskLabs framework, showcasing the potential of combining LLMs with traditional financial data for risk forecasting. By analyzing multiple data types — including earnings call transcripts, audio data, market time series, and contextual news — the study highlighted how LLMs can offer a more nuanced and multi-dimensional analysis of financial risks.

Moreover, Kim et al. [37] conducted a relevant study investigating the ability of ChatGPT-4 to perform financial statement analyses. They used balance sheets and income statements to let ChatGPT-4 predict either an increase or decrease in earnings in the next year. The results show that ChatGPT-4 can predict future earnings changes more accurately than human analysts, especially in complex scenarios where analysts typically struggle. These findings showed the potential of LLMs not only for traditional financial analysis but also for broader applications in risk assessment and auditing.

The capabilities of ChatGPT in auditing have also been explored by Eulerich et al. [28], who examined its performance on various accounting certification exams. Their study found that, with additional enhancements, ChatGPT successfully passed all sections of each tested exam. This supports the study by Kim et al. [37] that ChatGPT possesses the necessary capabilities to support auditing processes. Moreover, their findings suggest that optimizing ChatGPT can significantly improve its accuracy in domain-specific tasks.

In general, LLMs can be optimized through both parameter-tuning and prompt optimization, offering a more efficient way to tailor models to specific auditing tasks. Parameter-tuning refers to the fine-tuning of the model parameters, whereas prompt-tuning refines the instruction given to the model [53]. Another optimization technique is using few-shot learning, which describes showing the model multiple examples to help it generate the desired response [53]. Sarmah et al. [61] showcased the effectiveness of both prompt-tuning and few-shot learning by using a LLM with DSPy optimizers on various datasets. The DSPy framework[1] proposes a systematic approach to AI pipeline design by replacing manual prompt engineering with modular operators that automatically optimize language models [36]. Its optimizers use either prompt-tuning or few-shot learning, or a combination of both techniques. Sarmah et al. [61] found that the optimizers from the DSPy framework increase the performance of the LLM, highlighting the potential of programmatic prompt-tuning and few-shot learning to improve model accuracy.

In addition to the theoretical effectiveness of LLMs, it is crucial to consider their practical usability within auditing as well. AI adoption has been accepted and pursued in larger accounting firms, particularly among the Big Four [34]. However, most studies focused on the theoretical potential of LLMs, with relatively little attention given to their practical application in auditing [64]. A significant factor in this gap is the attitude of accountants toward integrating AI into their workflows. This was researched by Mulliqi [49], who noted that one of the primary challenges auditors face is "the fear of making mistakes, compounded by the lack of experimentation in a safe environment." Furthermore, the challenge of ensuring compliance with the standards for the use of AI tools adds another layer of complexity [49]. Given that their study addressed AI adoption broadly, it is essential to specifically investigate the practicalities of incorporating LLMs into auditing practices as well. This is also underlined by Kim et al. [37], who mentioned that further research is needed to determine whether LLMs can enhance human decision-making in practice.

Given the unique challenges of healthcare organizations described in previous sections, the ability of LLMs to process unstructured data, such as audit reports, becomes particularly valuable. By leveraging prompt-tuning and domain-specific adaptations, LLMs could enhance the audit in ways that traditional models cannot. This thesis tries to fill the gap in academic literature regarding the practical application of LLMs in auditing, with a special focus on the going concern assessment.

[1] DSPy is the second version of the earlier Demonstrate–Search–Predict (DSP) framework developed by Khattab et al. [35]. The framework is further discussed in Section 3.4.

Chapter 3

Method

This chapter describes the methodology used in this study. To answer the research question and validate the hypotheses outlined in Chapter 1, a design science approach was adopted to develop, evaluate, and refine the use of a large language model (LLM) to predict business continuity for healthcare organizations. The Design Science Research Process, as proposed by Peffers et al. [57], is well-suited for this research as it focuses on creating and evaluating models to solve real-world problems. By applying the design science framework, this study ensured that the developed LLM not only has technical effectiveness but also practical applicability within existing processes of the host company. Peffers et al. [57] formulated six activities as part of their Design Science Research Process, which are illustrated in Figure 3.1. The following section will describe how these activities have been implemented in this research.

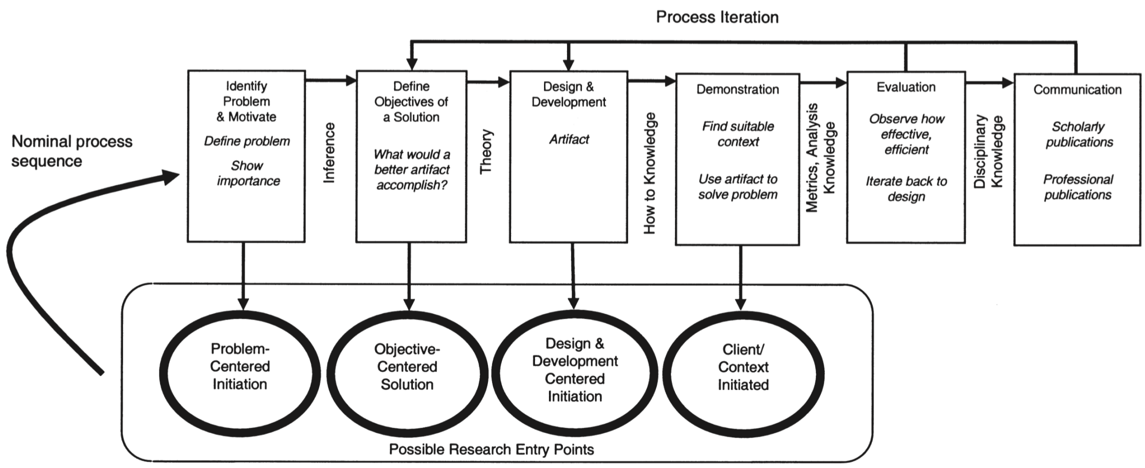


Figure 3.1: The activities of the Design Science Research Process by Peffers et al. [57].

Identify Problem & Motivate As explained in Chapter 1, the research problem centers on the challenges faced by auditors in predicting business continuity of healthcare organizations with financial uncertainty, particularly after the COVID-19 pandemic. Traditional machine learning methods have been used to enhance this process, but are limited in handling complex and high-volume data, while LLMs have not been applied extensively in this domain. Addressing this gap is essential as it could lead to more accurate business continuity assessments in auditing, thereby helping healthcare organizations to maintain operational stability and deliver critical public services.

Define Objectives of a Solution The primary objective of the solution was to develop a LLM-based model that can accurately predict the business continuity of healthcare organizations by analyzing both structured and unstructured data. Quantitative objectives included improving prediction accuracy over traditional methods and reducing the time required for assessments. The qualitative objectives focused on ensuring that the model is applicable within existing auditing processes and offering additional insights into the financial status of organizations in the healthcare sector.

These objectives have been captured in the following research question and hypotheses, as previously stated:

"How can LLMs be utilized and optimized effectively to enhance the going concern assessment for healthcare organizations?"

H.1 Programmatic prompt-tuned LLMs are more effective than traditional auditing methods in assessing the going concern.

H.2 Programmatic prompt-tuned LLMs can shorten the current process of assessing the going concern while maintaining accuracy.

H.3 Accountants perceive LLMs as a valuable tool to assist in going concern assessments.

Design & Development The artifact to be developed was a LLM-based model that was prompt-tuned to predict business continuity for healthcare organizations. The design involved defining the model's framework, capable of processing structured data and unstructured data. The development process included training and testing model optimizers on financial data and auditor's reports of healthcare organizations. The details of the development are discussed in the subsequent sections of this chapter.

Demonstration The demonstration of the model was done on the financial data of Dutch healthcare organizations and can be found in Chapter 4.

Evaluation We evaluated the effectiveness and efficiency of the LLM model through both qualitative and quantitative metrics. Quantitatively, the model's prediction accuracy was compared to traditional methods, while also assessing time savings. Additionally, the minimization of false negatives was important for the host company. In other words, they wanted to minimize the organizations that are predicted to not go bankrupt, even though they did. Qualitatively, feedback from accountants using the model in a real-world assessments was gathered to determine its ease of use and how well it fits into existing auditing processes. These metrics were measured by conducting a case study, which is discussed in Section 3.6. The interpretation of the evaluation can be found in Chapter 5. Subsequently, this led to recommendations to further optimize the model, which are outlined in Section 5.7.

Communication The findings, including the importance of the problem and the contributions to the academic and professional field, are summarized in Chapter 6.

The rest of this chapter discusses the various steps taken to create, optimize, test, and evaluate the LLM-based model, as part of the Design and Development activity.

3.1 Data collection and pre-processing

In line with other studies on bankruptcy prediction ([39] [12] [21] [73]), the main data included balance sheets and profit and loss statements, provided by BDO Netherlands. The dataset contained public data of healthcare organizations in the Netherlands, which are obligated to publish their yearly report (Wet marktordening gezondheidszorg, article 40b) [45]. In addition to the balance sheets and profit and loss statements, the data also included additional relevant information such as the type of healthcare organization, number of employees, number of patients, and various other features.

The provided dataset consisted of six separate Excel files, each corresponding to a different year from 2018 to 2023. For each file, the column names were cleaned and standardized, and several identifying columns were dropped (e.g. name and address). After the pre-processing steps, the cleaned Excel files were concatenated, resulting in a comprehensive dataset with one row per organization per year. It is important to note that not all columns were present in every file. In cases where a column was absent from a particular year's data, it was filled with NaN (Not a Number) values. Consequently, the final dataset used for modeling contains a significant number of NaN values across 340 columns.

To identify the organizations that have been declared bankrupt, the Dutch website "Faillissementen.com" was consulted. The time frame that was taken into account is 01-01-2018 until 16-10-2024, where only organizations with SBI code[1] 78, 82, 85, 86, 87, 88 are used, as these are the ones present in the dataset. The resulting list of bankrupt organizations was cross-referenced with our dataset using the unique

[1] In the Netherlands, the SBI code serves as a standard for categorizing economic activities, assigned by the Dutch Chamber of Commerce (KVK) [20].

identification number. A binary label was then added to each organization: '1' for bankrupt and '0' for non-bankrupt.

The auditor's reports used in the research were obtained from the website of the Dutch Ministry of Health, Welfare and Sport [22]. They are incorporated to test the LLM for its ability to handle unstructured data, as typically seen in these reports. Due to model and time constraints, only one page is taken for each organization. While this limitation reduces the volume of data analyzed, it still allows for the evaluation of the model's ability to handle unstructured data in general. The selection of the specific page was guided by input from accountants from the host company, who emphasized the importance of information not already captured in balance sheets or profit and loss statements. Therefore, we took the following process to select the page of the report:

1. Select the page with assets and obligations that are not mentioned in balance sheets (in Dutch: "Niet in de balans opgenomen activa en verplichtingen");
2. If not present, select a page containing financial metrics such as the liquidity ratio or the solvency ratio;
3. If both not present, select a page showing long- and/or short-term debts.

3.2 Feature selection

The initial dataset contained a substantial number of features, which were not all relevant to predict bankruptcy. Moreover, we were unable to consider the entire set of features, due to the high computational cost of the optimizers (see Section 3.4). To address this, feature selection was applied in two ways: expert opinion and mutual information scores. As previously mentioned, the focus is on minimizing false negatives, as requested by the host company. This is because it is unwanted to have many predictions for going concern, when the company did, in fact, go bankrupt. In addition to the prediction accuracy, this is what the choice of feature selection method was based on.

We began our training and testing process with 30 of the 340 features selected by accountants of the host company during interviews. These features were chosen based on the information they would use in the actual going concern assessment as part of the auditing process. The decision to include 30 features was arbitrary and based solely on the accountants' choice, rather than a formal selection criterion.

After the first trial with expert-based features, mutual information scores were calculated for every feature to see if this type of feature selection would improve the model performance. Mutual information is a measure of the dependency between features, where higher scores indicate stronger dependencies on the target feature [63]. Features with a higher mutual information score are then selected, as they convey more information with regards to the target feature. The calculation for mutual information scores cannot handle NaN values, so for numerical values, NaN values were imputed with the median and categorical values were imputed with 'Unknown'. However, the mutual information feature selection with 30 features did not result in a better model performance than the expert-based features.

We then proceeded to test the model with all features, but this worsened the prediction accuracy and false negatives as well. This meant that the features selected by the accountants performed the best, and were used for our analysis. While it is possible that another feature combination could result in a higher performance, identifying the optimal feature set was not the primary objective of this research.

Appendix A shows the features that were used in the model, in addition to the features that were the result of calculating the mutual information scores. The results of the comparison between the feature selection approaches can be seen in Chapter 4.

3.3 Sampling

For organizations that went bankrupt, we included the data from one or two years before bankruptcy. This is due to the limited sample size of organizations with data available only one year before bankruptcy (merely 6 organizations). This is also seen in other studies, where the time frame is usually one or two years before bankruptcy, or even up to six years before [1] [12]. This is discussed further in Section 3.5. For organizations that did not go bankrupt, all years were used for the model.

Subsequently, the dataset contained 41 organizations that have data from one or two years before bankruptcy. These were incorporated into the sample used to train and test the model. In addition, 41 random non-bankrupt organizations were added to create a balanced sample and to prevent biased predictions [38]. The total dataset used for training and testing was therefore 82 organizations. Since the sample was relatively small, we used stratified 3-fold cross-validation to maximize the use of available data and ensure robust model evaluation.

3.4 Model optimization with DSPy

The host company provided an API key for the use of OpenAI's GPT-4o model. This was therefore the baseline model to be refined and optimized. There were some limitations to the API key used to access this model:

- 20,000 tokens per minute
- 120 calls per minute

We used the DSPy framework by Khattab et al. [36] to programmatically optimize the model. There are several reasons to opt for the DSPy framework. Firstly, it enables you to create structured prompts with pre-defined input and output fields to ensure consistency across predictions. In addition, it can automate prompt optimization to further improve the performance of the model. Moreover, the framework has a modular approach to using LLMs, that allows for easier testing, debugging, and iterations of different aspects of the model. Lastly, DSPy handles the underlying API calls, allowing you to focus on the logic of the model rather than managing API interactions [36].

3.4.1 Aspects of the DSPy framework

The DSPy framework has a specific process with the following steps [65]:

1. Create examples
2. Define a signature
3. Use a module
4. Use an optimizer
5. Evaluate the output

(1). In DSPy, examples are the data used in the model, to be divided into train and test sets [25]. In our case, an example is one organization with financial data for a specific year, as well as the path to the PNG of the auditor's report of that year.

(2). Signatures are used to break the input and output of the model into abstract parts. It is a "natural-language typed declaration of a function" [36], that tells the model what the text transformation should be instead of what the prompt should be. This is done by taking the examples and defining the input and output fields that DSPy then parses into meaningful instructions for the LLM [36]. Our signature is defined as follows:

Table 3.1: Definition of our Signature within DSPy.

Type	**Name**	**Description**
Input	`financial_data`	`dspy.InputField()`
Input	`png_summary`	`dspy.InputField(default="")`
Output	`bankruptcy_status`	`dspy.OutputField(desc="Gegeven de reasoning, reageer uitsluitend met 'ja' als de organisatie binnen twee jaar failliet zal gaan; reageer anders met 'nee'.")` [2]
Output	`explanation`	`dspy.OutputField(desc="Geef een korte (nederlandse) uitleg voor de voorspelling door de belangrijkste financiële gegevens te vermelden die je hebt gebruikt.")` [3]

The output of the model was used for the case study, performed by Dutch accountants. Therefore, the prompts and output of the model were in Dutch to ensure their understanding of the predictions.

(3). The next step is to declare the module that describes how the LLM should be used. The default module, Predict, is fundamental to all other modules and primarily enables the LLM to predict what the output should be [55]. To compare, the module ChainOfThought prompts the LLM to perform a reasoning process before giving its final response [55]. Since the response of the model should be the result of a logical thought process, the ChainOfThought module is used in our model. Within our prediction module, we do an additional step by first giving the PNG of the auditor's report as input to the model, which returns a summary of the image. This is a separate API call, and the output is used as the input field in the official prediction module to be optimized. The following prompt was used to generate the summary:

"Analyze this financial image and summarize the key financial information visible. Focus on indicators that may suggest the company's financial health or a risk of bankruptcy, and keep it concise with a maximum of 100 tokens. Ensure that the organization's name is not included in your response." (Translated from Dutch)

The limitation of a 100 tokens was also defined in the API call, meaning that whenever the model would go over the limit, the summary would be cut off at 100 tokens to ensure a concise and focused input for the model later in the process. Following this initial API call to generate the auditor's report summary, the actual prediction model begins, using the previously defined Signature.

(4). The optimization step incorporates various optimizers, previously referred to as "teleprompters" [26], to enhance the model's performance. This is explained in the next section.

(5). The performance of the optimized model is assessed using an evaluation function. The function compares the model's output—the predicted bankruptcy status—against the actual label, returning a binary True or False result. The explanation of the prediction, which is the second part of the output, is not evaluated within this evaluation function and therefore not taken into account for optimization. This decision was made primarily due to the complexity of automatically assessing the quality of natural language explanations. Evaluating explanations would require sophisticated natural language processing techniques or human judgment, which are beyond the scope of this research. Instead, the focus is on the model's primary task of correctly predicting bankruptcy status, which can be objectively measured and optimized. However, Chapter 5 does discuss the difference between the explanations of the model and the accountants on a qualitative level.

Figure 3.2 visualizes the process that the model follows within our code. The process begins with the `optimize_and_test` function, which initiates the optimizer and proceeds to the training phase. Key components include the `DynamicPredictionModule` and `DynamicBankruptcyPrompt` classes, which define the module and signature, respectively. The forward function initiates the prediction process, utilizing `ChainOfThought` reasoning and the `openAI_PROCESS_PNG` function for the summarization of the auditor's report page. The model goes through iterative optimization, adjusting prompts and examples based on the optimizer used. Finally, the predictions are evaluated in the `evaluate_model` function, and the process stops when both the training and testing phase are complete.

3.4.2 Optimizers

DSPy offers various optimizers to fine-tune the model in order to optimize its performance. In this research, we chose a variety of these optimizers to see their effect on accuracy. The decision was based on the inclusion of different types of optimizers, as well as their presence in the literature [61] [37] [44]. Table 3.2 summarizes the DSPy optimizers used in this research, which are further explained in the following section. The descriptions are based on information from DSPy's official website and its source code [27] [54] [55].

Labeledfewshot This first optimizer provides the model with k randomly-chosen labeled examples from the train set to assist in its prediction for the test set. These examples are therefore organizations that the model has not seen before.

[2]Translation: "Given the reasoning, respond with only 'yes' if the organization will go bankrupt within two years; otherwise, respond 'no'."

[3]Translation: "Provide a brief (Dutch) explanation for the prediction by listing the key financial data you used."

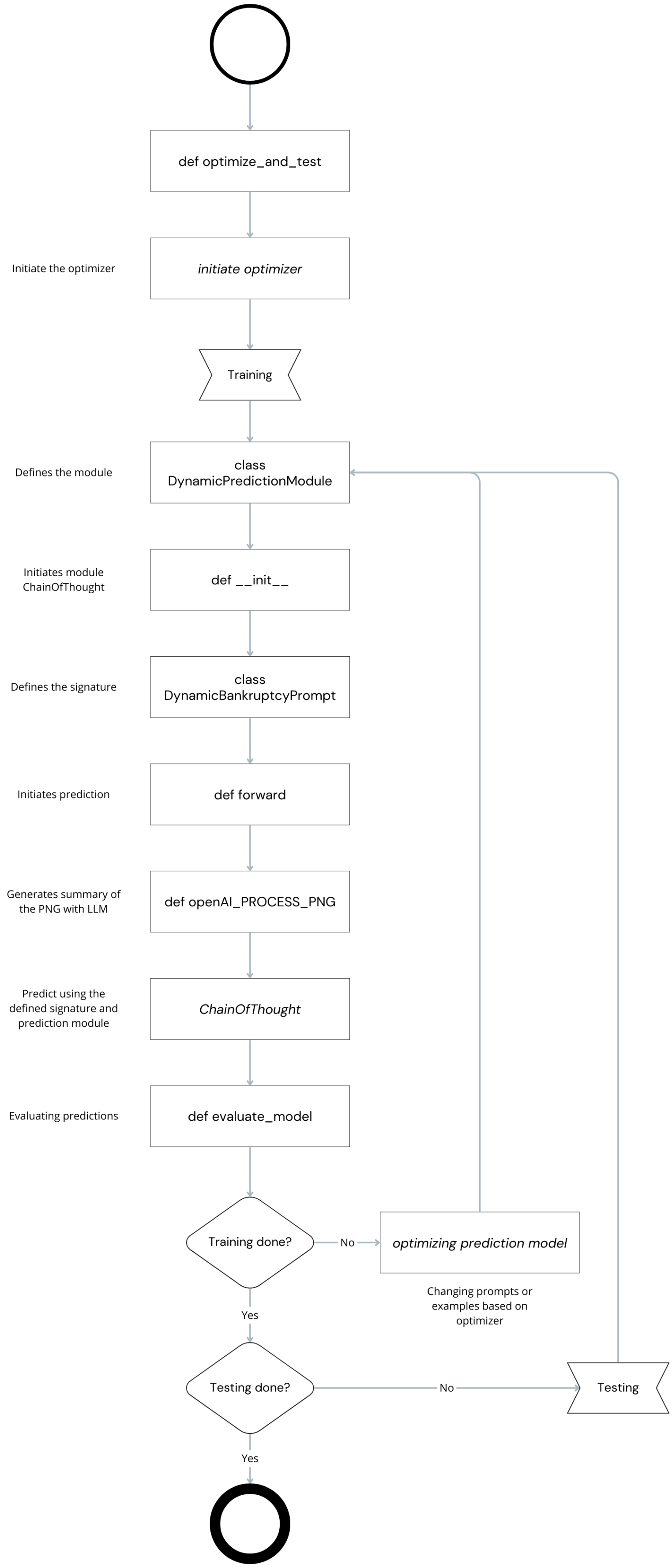


Figure 3.2: Flow chart of the code of our model.

Table 3.2: Overview of DSPy Optimizers used in this study.

Optimizer	Few-shot Learning	Prompt Optimization	Search Strategy
LabeledFewShot	Yes	No	Randomized selection
BootstrapFewShot	Yes	No	Greedy selection
BootstrapFewShotWith RandomSearch	Yes	No	Randomized selection
KNNFewShot	Yes	No	Nearest neighbors
COPRO	No	Yes	Iterative refinement
MIPROv2	Yes	Yes	Hybrid approach

BootstrapFewShot The BootstrapFewshot optimizer operates as follows:

1. The train set is used to let the model predict bankruptcy. The organizations that the model predicts correctly are given as an example for the test set.
2. It stops searching once it finds k desired examples for the prompt (i.e. the maximum amount of bootstrapped examples).
3. In addition to the bootstrapped examples, it also gives a pre-defined number of labeled examples to the test set.

BoostrapFewShotWithRandomSearch While BootstrapFewShot selects the first k examples, BootstrapFewShotWithRandomSearch has a more extensive approach for selecting examples. For each seed, it generates a model differently:

1. Seed = -3 (Zero-shot):
   - Lets the model predict without examples.
2. Seed = -2 (Labels only):
   - Uses LabeledFewShot with maximum number of labeled examples.
3. Seed = -1 (Unshuffled few-shot):
   - Applies BootstrapFewShot without shuffling the train set.
4. Seeds ≥ 0 (Random search):
   - Shuffles the train set using the current seed.
   - Randomly selects a number of examples between the minimum number of examples and maximum number of examples.
   - Applies BootstrapFewShot with these randomly selected examples.

After iterating through all the seeds, the best performing model is taken to predict for the test set.

KNNFewShot With the KNNFewShot optimizer, the following process is repeated for each example in the test set:

1. The optimizer finds the k nearest neighbors for each example by calculating the cosine similarity between the input example and all examples in the train set.
2. The k examples with the highest similarity scores are selected as the nearest neighbors.
3. Once the nearest neighbors are identified, the BootstrapFewShot process is applied:
   (a) The optimizer gives a prediction for each of the k examples, until either the maximum number of bootstrapped demos is reached or all k examples have been processed.
   (b) The successful predictions are taken as examples for the example of the test set.

COPRO The COPRO optimizer focuses on optimizing the prompt, instead of providing examples to improve the model. It does so in the following matter:

1. The LLM generates new instruction candidates (breadth=3), i.e. different prompts, that would optimize bankruptcy prediction.
2. These instruction candidates are evaluated on the whole train set and their accuracy scores are documented.
3. The candidates and their scores are then fed back into the LLM, which will be asked to generate new prompts based on the evaluation of the previous ones. This is done a certain number of iterations (depth=3).
4. The best candidate of all iterations is used on the test set.

MIPROv2 Similar to COPRO, the MIPROv2 optimizer adjusts the prompt, but additionally applies few-shot learning by using examples for the test set. The validation set is defaulted at 80% of the train set. The MIPROv2 optimizer has the following process:

1. The optimizer begins with a bootstrapping stage using the BootstrapFewShotWithRandomSearch approach.
2. Then, the LLM generates multiple instruction candidates (i.e. prompts) considering the following data:
    - Properties of the train dataset;
    - The LLM program's code and the predictor that the instruction is made for;
    - The previously bootstrapped examples with inputs and outputs;
    - A randomly sampled hint for the generation (e.g. to be creative or concise).
3. Next, MIPROv2 initiates a search stage, where this process is executed for a specified number of trials:
    (a) It samples mini-batches from the train set.
    (b) It proposes combinations of instructions and examples that are tested on these mini-batches.
4. The best-performing combination of instructions and examples is selected for use on the test set.

Table 3.3 describes the hyperparameters used in the DSPy optimizers. For most hyperparameters, a number lower than the default is chosen because of the time limitation. They were, however, the same for all optimizers to make the comparison as fair as possible. The number of threads refers to the number of parallel calls made to the model. This was kept at 1 to keep the output clear and straightforward.

3.5 Boundaries of the model

As discussed in Section 3.1, only data from one and two years prior to bankruptcy was considered in our model. The sample of bankrupt companies with data from one year before bankruptcy was too small, consisting of only six companies. Therefore, data from two years before bankruptcy was included as well. This approach aligns with the findings of Bellovary et al. [12], who conducted a literature review of bankruptcy studies spanning from the 1930s to 2007. Their study shows that some studies even use data from as far back as six years before bankruptcy. However, due to the computational expenses of the optimizers used in our model, incorporating this extended time frame was not feasible within our research. Nevertheless, to explore the impact of the time frame, we conducted a small-scale test using a subset of organizations with data from multiple years before bankruptcy to assess its effect on the baseline model.

For this analysis, we created a subset of 20 organizations for each year, ranging from two to six years before bankruptcy. To ensure balanced representation and mitigate prediction bias, we selected 10 bankrupt and 10 non-bankrupt organizations for each year. We used the baseline model to test the boundaries of the model, where for every year before bankruptcy, the whole dataset of 20 organizations is used for testing. This is because there is no training involved with the baseline model and the samples are relatively small, and we wanted to prevent variability from impacting the results. The results of this temporal analysis are presented in Section 4.2.

Table 3.3: Hyperparameter settings for the DSPy optimizers [54].

DSPy Optimizer	**Hyperparameters**
Baseline	None
Labeled Fewshot	k=6
Bootstrap Few Shot	max_bootstrapped_demos=3 (default=4) max_labeled_demos=6 (default=16)
Bootstrap Few Shot Random Search	max_bootstrapped_demos=3 (default=4) max_labeled_demos=6 (default=16) num_candidate_programs=3 (default=16) num_threads=1 (default=6)
KNN Few Shot	k=6
COPRO	max_bootstrapped_demos=3 (default=4) max_labeled_demos=6 (default=16) num_threads=1 (default=unspecified) depth=3 (default=3) breadth=3 (default=10)
MIPROv2	max_bootstrapped_demos=3 (default=4) max_labeled_demos=6 (default=16) num_threads=1 (default=6) num_candidates=3 (default=10) num_trials=10 (default=30) minibatch_size=10 (default=25)

3.6 Case Study

We conducted a case study with accountants to evaluate the effectiveness of the LLM model in assessing going concern both qualitatively and quantitatively. Although the setup resembles a controlled experiment in its structure and comparison between two groups, we refer to it as a case study throughout this thesis. This is due to the fact that the focus was not on strict experimental control, but rather on evaluating the model's practical use in a structured setting.

The study aimed to compare the model's performance to traditional auditing methods, focusing on three key aspects:

- Prediction accuracy - how well the predictions align with actual bankruptcy outcomes.
- Time efficiency - the extent to which the model reduces the time required for the going concern assessment.
- Satisfaction with the model - the satisfaction with the performance of the model and the ease with which accountants can incorporate the model's predictions into their workflow.

To this end, the model's predictions were compared to the predictions of accountants. The case study was designed as a controlled experiment, where accountants assessed a subset of healthcare organizations using either the LLM-assisted approach or traditional methods. It is important to note that the study was not intended to replicate a real-world audit, but to measure the practicality and effectiveness of the model in a structured setting.

The model predictions used in the case study were generated by the best-performing optimizer, as determined in Section 4.1. The organizations included in the case study were selected from the test set of the first fold, as its confusion matrix most closely resembled the aggregated confusion matrix (see Section 4.1) and is therefore a good representation. Since the test set contained 28 organizations and the case study only required 20, a random selection is made of those 28, while maintaining the proportional distribution of the aggregated confusion matrix.

The case study was conducted with two groups of accountants, each consisting of five participants. Table 3.4 provides an overview of the characteristics of the participants. All participants were employed at BDO and are (or were previously) accountants within the organization. To be able to make a valid comparison, the participants were divided between the two groups based on their role within BDO, as

Table 3.4: Overview of participants of the case study.

Participant	Role	RA Title	Department	Group
Participant 1a	Junior manager	No	Audit & Assurance	Control
Participant 1b	Junior manager	No	Audit & Assurance	Control
Participant 1c	Equity partner	Yes	Audit & Assurance	Control
Participant 1d	Senior manager	Yes	Audit & Assurance	Control
Participant 1e	Manager	Yes	IT Audit (previously Audit & Assurance)	Control
Participant 2a	Salary partner	Yes	Audit & Assurance	LLM
Participant 2b	Senior manager	Yes	IT Audit (previously Audit & Assurance)	LLM
Participant 2c	Manager	Yes	Audit & Assurance	LLM
Participant 2d	Junior manager	No	Audit & Assurance	LLM
Participant 2e	Manager	Yes	Audit & Assurance	LLM

well as possession of the RA title. RA stands for "Register Accountant" in Dutch, which is the official title for accountants registered with the Dutch professional organization of accountants (NBA) [68].

To be able to see the difference between the performance of accountants with the model predictions and their performance without, the two groups received different types of data:

1. Control group - received only financial data and one page of the auditor's report.
2. LLM group - also received financial data and one page of the auditor's report, but additionally got the prediction of the model and a model-generated summary of the auditor's report.

Both groups had 45 minutes to evaluate 20 organizations and determined whether each would go bankrupt within two years based on the available data. After 45 minutes, the participants were told to stop their assessment, regardless of their progress with the evaluation of the organizations. The participants completed the assessment individually while being supervised and did not use any external tools like a calculator or the internet.

At the end of the session, the accountants filled in a feedback form with several evaluation questions. Both groups were asked to reflect on how they experienced the assessment process and if they felt any relevant financial information was missing. In addition, the LLM group received further questions aimed at understanding their interaction with the model. These included how much they relied on the model's predictions, which aspects they found most valuable or insightful, whether they would consider using such a model in real-world assessments, and any further feedback they had on the tool.

Examples of the case study materials and evaluation form for both groups can be found in Appendix C.

Chapter 4

Results

This chapter presents the results of the model and the case study, as described in the methodology. We first discuss the accuracy of the model, after which we present the results of the temporal analysis to test the boundaries of the model. Subsequently, the case study results are outlined in four separate sections: speed, accuracy, ease of assessment, and satisfaction with the model.

4.1 Model accuracy

Tables 4.1 to 4.4 present the evaluation metrics for the different optimizers used in the model. These metrics include accuracy, F1-score, and false negatives, with a particular focus on minimizing false negatives, as requested by the host company. The metrics are calculated per fold of the 3-fold cross-validation, as well as for the aggregated results, which serve as the basis for comparing the optimizers. The full table of model results is presented in Appendix B.

Our initial step was to analyze the impact of feature selection on model performance. As shown in Table 4.4, the baseline model scores best when using only the features selected by accountants, with an accuracy of 78%. In contrast, selecting a subset of the features based on mutual information scores led to a lower accuracy (76%) and an increase in false negatives (14 vs. 12). Furthermore, testing the baseline model with all features resulted in a further decrease in accuracy (73%). For this reason, the baseline model with expert-based features was chosen as the foundation for optimization.

Building on this, we then evaluated the performance of the optimizers that used the baseline model with expert-based features. Examining the aggregated accuracy, all optimizers scored within a range of 76% - 79%. Notably, not all optimizers led to improvements over the baseline, which already achieved 78% accuracy with expert-based features. BootstrapFewShotWithRandomSearch and KNN-FewShot both underperformed compared to the baseline, showing lower accuracy and similar false negative rates. Meanwhile, BootstrapFewShot and MIPROv2 achieved accuracy levels comparable to the baseline, though BootstrapFewShot had the highest number of false negatives among all optimizers (16). Only the LabeledFewShot and COPRO optimizers outperformed the baseline in terms of accuracy with 79%. In comparison, the COPRO optimizer produces fewer false negatives than the LabeledFewShot optimizer (10 vs. 13). Given that the host company prioritized the minimization of false negatives alongside achieving high accuracy, the COPRO optimizer is the best-performing option of all optimizers. Nevertheless, it is important to emphasize that the performance differences are minimal and should be interpreted with caution.

Table 4.1: Results of optimizers for Fold 1.

Optimizer	Fold 1			
	Accuracy	F1	CM	FN
Baseline (all columns)	79%	0.78	[[10 4] [2 12]]	4
Baseline MI-based features	79%	0.78	[[10 4] [2 12]]	4
Baseline expert-based features	82%	0.82	[[11 3] [2 12]]	3
LabeledFewShot	79%	0.78	[[10 4] [2 12]]	4
BootstrapFewShot	75%	0.75	[[9 5] [2 12]]	5
BootstrapFewShotWithRandomSearch	82%	0.82	[[11 3] [2 12]]	3
KNNFewShot	82%	0.82	[[11 3] [2 12]]	3
COPRO	82%	0.82	[[11 3] [2 12]]	3
MIPROv2	79%	0.78	[[10 4] [2 12]]	4

Table 4.2: Results of optimizers for Fold 2.

Optimizer	Fold 2			
	Accuracy	F1	CM	FN
Baseline (all columns)	63%	0.62	[[7 7] [3 10]]	7
Baseline MI-based features	74%	0.74	[[9 5] [2 11]]	5
Baseline expert-based features	74%	0.73	[[8 6] [1 12]]	6
LabeledFewShot	74%	0.73	[[8 6] [1 12]]	6
BootstrapFewShot	74%	0.73	[[7 7] [0 13]]	7
BootstrapFewShotWithRandomSearch	67%	0.66	[[7 7] [2 11]]	7
KNNFewShot	67%	0.66	[[8 6] [3 10]]	6
COPRO	74%	0.74	[[9 5] [2 11]]	5
MIPROv2	70%	0.69	[[7 7] [1 12]]	7

Table 4.3: Results of optimizers for Fold 3.

Optimizer	Fold 3			
	Accuracy	F1	CM	FN
Baseline (all columns)	78%	0.78	[[10 3] [3 11]]	3
Baseline MI-based features	74%	0.74	[[8 5] [2 12]]	5
Baseline expert-based features	78%	0.78	[[10 3] [3 11]]	3
LabeledFewShot	85%	0.85	[[10 3] [1 13]]	3
BootstrapFewShot	85%	0.85	[[9 4] [0 14]]	4
BootstrapFewShotWithRandomSearch	81%	0.81	[[11 2] [3 11]]	2
KNNFewShot	78%	0.78	[[9 4] [2 12]]	4
COPRO	81%	0.81	[[11 2] [3 11]]	2
MIPROv2	85%	0.85	[[11 2] [2 12]]	2

Table 4.4: Aggregated results of optimizers (run time measured in hh:mm).

Optimizer	Aggregated				
	Run time	Accuracy	F1	CM	FN
Baseline (all columns)	00:23	73%	0.73	[[27 14] [8 33]]	14
Baseline MI-based features	00:15	76%	0.75	[[27 14] [6 35]]	14
Baseline expert-based features	00:13	78%	0.78	[[29 12] [6 35]]	12
LabeledFewShot	00:31	79%	0.79	[[28 13] [4 37]]	13
BootstrapFewShot	00:35	78%	0.77	[[25 16] [2 39]]	16
BootstrapFewShot WithRandomSearch	01:26	77%	0.76	[[29 12] [7 34]]	12
KNNFewShot	02:14	76%	0.75	[[28 13] [7 34]]	13
COPRO	**05:39**	**79%**	**0.79**	**[[31 10] [7 34]]**	**10**
MIPROv2	03:25	78%	0.78	[[28 13] [5 36]]	13

Taking a closer look at the COPRO optimizer, Tables 4.5, 4.6, and 4.7 present the prompts generated by the optimizer along with their corresponding accuracy scores, divided into a table per fold. The COPRO optimizer generates an initial set of three prompts (breadth = 3) and evaluates each prompt on every sample in the train set to obtain accuracy scores. This constitutes the first iteration (depth = 1). The accuracy scores, along with the corresponding prompts, are then fed into the LLM, which generates three new prompts based on this evaluation. This process is repeated twice, reaching a maximum depth of three. At this point, the best-performing prompt is selected and applied to the test set. This prompt is highlighted in each table.

The accuracy scores for the prompts range from 72.2% to 83.6%, similar to the lowest and highest accuracies observed for the optimizer when considering individual folds. Within each fold, the prompts are grouped by iteration depth. As shown in the tables, the accuracy scores in folds 1 and 2 exhibit a slight increase in later iterations compared to the initial prompts. However, in the third fold, the scores decline across iterations. The optimization in fold 3 resulted in two prompts achieving the highest score, with the prompt from the first iteration ultimately selected by the optimizer to use on the test set.

Table 4.5: The evaluation scores of the COPRO-generated prompts - Fold 1.

	Prompt	Score
1	Analyze the provided financial data and png_summary to determine the bankruptcy_status of the entity in question. Additionally, provide a detailed explanation describing your reasoning behind the status determination, utilizing specific points and data from the given information.	72.2%
2	Analyze the financial_data and generate a visual png_summary scrutinizing the data. Based on this analysis, predict the bankruptcy_status of the entity and provide a comprehensive explanation of the reasoning behind your prediction.	72.2%
3	Given the fields financial_data, png_summary, produce the fields bankruptcy_status, explanation.	77.8%
4	**Evaluate the given financial_data and generate a comprehensive png_summary visualizing the key aspects of the data. Using your analysis of these visual representatives, determine the bankruptcy_status of the entity. Provide a detailed explanation clearly outlining the data-driven reasoning behind your bankruptcy status prediction.**	**79.6**%
5	Evaluate the given financial_data and analyze the png_summary to accurately predict the bankruptcy_status of the entity. Provide a well-elaborated explanation for your determination, making sure to integrate relevant financial indicators and notable insights visualized in the graphs.	75.9%
6	Analyze the provided financial_data and review the given png_summary. Using this information, determine the bankruptcy_status of the entity and write a detailed explanation that explains the reasoning behind your status determination. Ensure that your explanation clearly references specific data points and trends from the financial analysis and visual summary.	77.8%
7	Assess the provided financial_data, and scrutinize the information in the png_summary thoroughly. Utilize this detailed examination to ascertain the bankruptcy_status of the entity and compose an in-depth explanation discussing the rationale behind your determination. Highlight the critical financial indicators and signal trends elucidated in your visual and numerical analysis, ensuring each critical point is meticulously addressed.	74.1%
8	Using the given financial_data and the extracted png_summary, evaluate and conclude the entity's bankruptcy_status. Provide a thorough explanation articulating the reasoning behind your conclusion, anchoring it in salient data points and identified trends.	75.9%
9	Analyze the provided financial_data and interpret the corresponding png_summary that visualizes key trends from the data. Determine the company's bankruptcy_status based on a thorough analysis and construct a detailed explanation supporting your status determination. Make sure your explanation integrates specific references from both financial data and visual summaries, often studying specific metrics and data point changes over time.	77.8%

Table 4.6: The evaluation scores of the COPRO-generated prompts - Fold 2.

	Prompt	Score
1	Analyze the provided financial_data and png_summary to determine the bankruptcy_status of the entity in question. Additionally, provide a detailed explanation describing your reasoning behind the status determination, utilizing specific points and data from the given information.	78.2%
2	Analyze the financial_data and generate a visual png_summary scrutinizing the data. Based on this analysis, predict the bankruptcy_status of the entity and provide a comprehensive explanation of the reasoning behind your prediction.	81.8%
3	Given the fields financial_data, png_summary, produce the fields bankruptcy_status, explanation.	78.2%
4	Analyze the provided financial_data to generate a visual png_summary that clearly highlights the key aspects of the data. Based on this visual summary, assess the financial stability of the entity and determine its bankruptcy_status. Thoroughly support your assessment with a detailed explanation, referring explicitly to the most significant points and data from the summary to justify your conclusion.	81.8%
5	Review the given financial_data and png_summary to estimate the bankruptcy_status of the subject entity. Subsequently, craft a thorough explanation detailing the rationale behind your conclusion, ensuring to highlight pertinent data and trends observed in the provided information for clarity and support.	78.2%
6	Examine the given financial_data in conjunction with the png_summary to predict the entity's bankruptcy_status. Provide a logical and well-structured explanation that comprehensively justifies your prediction, taking into account specific details and data points extracted from the provided documents.	81.8%
7	**Carefully examine the financial_data provided to derive insights about the entity's financial health. Next, create a clear and concise png_summary that encapsulates the critical financial aspects visually. Use this visual summary to predict whether the entity is likely to go bankrupt, and then elaborate with an in-depth explanation that outlines your criteria and reasoning, citing specific details and data points from both the financial data and the visual summary.**	**83.6%**
8	Analyze the provided financial_data in detail and create a comprehensive png_summary highlighting all key financial metrics. Use both the raw numerical data and the visual png_summary to thoroughly evaluate the entity's overall financial health and predict its bankruptcy_status. Provide a logically organized and data-driven explanation for your prediction, citing concrete figures and visual indicators to support your conclusion.	80.0%
9	Thoroughly analyze the given financial_data and use it to create a comprehensive visual png_summary. Next, examine both the original data and the visual summary meticulously to evaluate the financial health of the entity, determine the likelihood or bankruptcy_status, and explicate your findings in a detailed explanation. The explanation should include reasoned arguments referencing specific data points and features illustrated in the summary to bolster your well-reasoned prediction.	81.8%

Table 4.7: The evaluation scores of the COPRO-generated prompts - Fold 3.

	Prompt	Score
1	**Analyze the provided financial_data and png_summary to determine the bankruptcy_status of the entity in question. Additionally, provide a detailed explanation describing your reasoning behind the status determination, utilizing specific points and data from the given information.**	**81.8%**
2	Analyze the financial_data and generate a visual png_summary scrutinizing the data. Based on this analysis, predict the bankruptcy_status of the entity and provide a comprehensive explanation of the reasoning behind your prediction.	80.0%
3	Given the fields financial_data, png_summary, produce the fields bankruptcy_status, explanation.	76.4%
4	Thoroughly evaluate the provided financial_data and the png_summary. Based on your analysis, determine the bankruptcy_status of the entity. Follow your determination with a detailed and fact-based explanation, explicitly referencing relevant data to support your reasoning.	81.8%
5	Using the provided financial_data and the png_summary, conduct a thorough analysis to assess the bankruptcy_status of the entity.	80.0%
6	Review the given financial_data and png_summary. Based on this review, swiftly determine the bankruptcy_status of the entity and provide a thorough explanation. Your explanation should clearly delineate each step taken in reaching the bankruptcy status assessment, ensuring to integrate significant data points and highlights from both the financial_data and png_summary.	70.9%
7	Evaluate the given financial_data and the png_summary to ascertain the bankruptcy_status of the entity under consideration. Offer a thorough explanation highlighting your reasoning, referencing specific data points and insights from both the financial data and the visual summary to substantiate your conclusion.	80.0%
8	Examine the available financial_data and provided png_summary. Your task is to predict the bankruptcy_status of the entity. Provide an in-depth explanation, highlighting key financial indicators, patterns, and relevant visual data points that led to your conclusion.	72.7%
9	Carefully evaluate the given financial_data and any relevant figures in the png_summary. Determine the entity's bankruptcy_status and frame a comprehensive explanation noting the critical data and trends. Draw on concrete metrics and visual insights to reinforce your rationale, ensuring that each inference ties back clearly to the provided financial details.	74.5%

4.2 Boundaries of the model

As described in the methodology, we tested the boundaries of the model's capabilities by analyzing how far into the future it could accurately predict bankruptcy. This analysis involved using data from 2 to 6 years prior to bankruptcy, while using the baseline model with expert-based features for bankruptcy prediction. The results of this temporal analysis are presented in Table 4.8. Since the baseline model does not involve a training phase, the whole dataset was used for testing without applying k-fold cross-validation. As a result, only the evaluation metrics for the complete dataset are presented, without any individual folds.

The results show a downward trend in accuracy as the prediction time frame increases. The model achieves relatively accurate predictions up to 3 years before bankruptcy, after which accuracy declines. In particular, when predicting five years before bankruptcy, the accuracy of the model falls below 50%.

Table 4.8: The results of the temporal analysis (run time measured in hh:mm).

Years before bankruptcy	**Run time**	**Accuracy**	**F1**	**CM**	**FN**
2 years before bankruptcy	00:03	75%	0.73	[[5 5] [0 10]]	5
3 years before bankruptcy	00:04	85%	0.85	[[7 3] [0 10]]	3
4 years before bankruptcy	00:03	55%	0.52	[[3 7] [2 8]]	7
5 years before bankruptcy	00:04	45%	0.45	[[4 6] [5 5]]	6
6 years before bankruptcy	00:03	65%	0.63	[[4 6] [1 9]]	6

4.3 Case study

Table 4.9 and Table 4.10 show the results of the case study conducted with accountants. In the study, accountants were tasked with predicting whether an organization would go bankrupt within two years, providing a binary response of either 'Yes' or 'No', as well as an explanation for their response. The control group received just the financial data and the auditor's report page, while the LLM group also received the model's prediction and a model-generated summary of the auditor's report page.

4.3.1 Accuracy

When comparing the accuracy of the individual accountants with the model, it becomes evident that the model predicts the same or better than every single accountant with an accuracy of 85%. In other words, the model matches the performance of the two highest-scoring accountants in the case study, while scoring better than the others. Secondly, when considering the average accuracy, the control group and the LLM group exhibit very similar results, with accuracies of 77% and 78%, respectively.

Since the average results do not indicate any difference between the groups, it is worth looking deeper into the individual organizations that were evaluated. Many organizations have received the same evaluation from both the model and the accountants. We can derive that these were clear-cut cases where bankruptcy (or non-bankruptcy) was evident. Focusing on the cases where accountants disagreed on a prediction, as shown in Table 4.11, we observe that the LLM group achieved slightly higher accuracy (67% vs. 60%).

4.3.2 Speed

In addition to accuracy, we measured the speed with which the accountants were able to evaluate the organizations. The accountants were given a 45-minute time frame to conduct their evaluations, allowing us to observe how many organizations they would evaluate within that period.

The results of the control group show that three of the five participants managed to evaluate all 20 organizations, while the other two were able to assess 13 and 19. Similar to the control group, three participants of the LLM group evaluated all 20 organizations, whereas the other two evaluated 15 and 17. As a result, the average number of organizations evaluated per participant was 18.4 for both groups.

Table 4.9: The results of the control group of the case study.

Organization	**Actual**	**Model prediction**	**1a**	**1b**	**1c**	**1d**	**1e**	**Average**
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
2	No	No	No	No	No	No	No	
3	Yes	Yes	No	No	No	No	Yes	
4	No	No	No	No	No	No	No	
5	Yes	No	No	No	Yes	Yes	No	
6	No	No	No	No	No	No	No	
7	No	Yes	Yes	No	Yes	Yes	No	
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
9	No	No	No	No	Yes	No	No	
10	No	No	No	No	No	No	No	
11	Yes	No	No	No	No	No	No	
12	Yes	Yes	No	Yes	Yes	Yes	Yes	
13	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
14	No	No	No	-	No	No	No	
15	Yes	Yes	No	-	No	No	No	
16	No	No	No	-	No	No	No	
17	Yes	Yes	Yes	-	Yes	Yes	Yes	
18	No	No	No	-	No	No	No	
19	No	No	No	-	No	No	No	
20	Yes	Yes	-	-	Yes	Yes	Yes	
Accuracy		85%	68%	77%	75%	80%	85%	**77%**
Nr. Organizations		20	19	13	20	20	20	**18.4**

Table 4.10: The results of the LLM group of the case study.

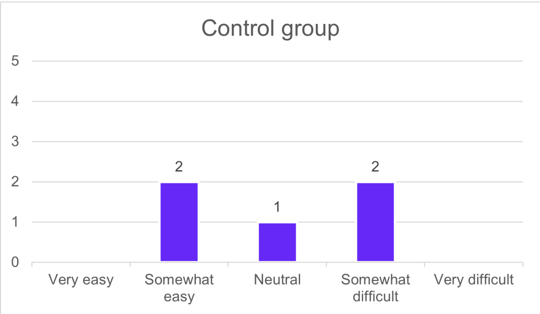
Organization	**Actual**	**Model prediction**	**2a**	**2b**	**2c**	**2d**	**2e**	**Average**
1	Yes	Yes	Yes	Yes	Yes	Yes	No	
2	No	No	No	No	No	No	No	
3	Yes	Yes	Yes	No	Yes	No	Yes	
4	No	No	No	No	No	No	No	
5	Yes	No	No	No	Yes	No	No	
6	No	No	No	No	No	No	No	
7	No	Yes	Yes	Yes	Yes	No	No	
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
9	No	No	No	No	No	No	No	
10	No	No	No	No	No	No	No	
11	Yes	No	No	No	No	No	No	
12	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
13	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
14	No	No	No	No	No	No	No	
15	Yes	Yes	No	No	No	No	No	
16	No	No	No	No	No	-	No	
17	Yes	Yes	Yes	Yes	Yes	-	Yes	
18	No	No	-	No	No	-	No	
19	No	No	-	No	No	-	No	
20	Yes	Yes	-	Yes	Yes	-	Yes	
Accuracy		85%	76%	75%	85%	73%	80%	**78%**
Nr. Organizations		20	17	20	20	15	20	**18.4**

Table 4.11: Results of the edge cases within the case study of control group (1) and LLM group (2).

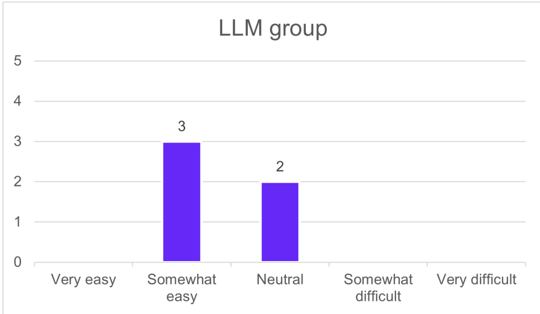
Org.	**Actual**	**1a**	**1b**	**1c**	**1d**	**1e**	**Average**	**2a**	**2b**	**2c**	**2d**	**2e**	**Average**
1	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	No	
3	Yes	No	No	No	No	Yes		Yes	No	Yes	No	Yes	
5	Yes	No	No	Yes	Yes	No		No	No	Yes	No	No	
7	No	Yes	No	Yes	Yes	No		Yes	Yes	Yes	No	No	
9	No	No	No	Yes	No	No		No	No	No	No	No	
12	Yes	No	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes	
	Accuracy	33%	67%	50%	67%	83%	**60%**	67%	50%	83%	67%	67%	**67%**

4.3.3 Ease of assessment

At the end of the case study, the groups filled in an evaluation form. The full answers can be seen in Appendix C. The accountants were asked to describe the ease with which they were able to assess bankruptcy in the case study. As seen in Figures 4.1a and 4.1b, none of the participants found it particularly easy or difficult. However, the control group thought the assessment was slightly more difficult than the LLM group.



(a) Control group.



(b) LLM group.

Figure 4.1: Comparison of ease of assessment between the control group and the LLM group.

Appendix C presents the explanations provided by the accountants regarding the difficulty of the case study. A key trend observed in their responses was the perceived lack of certain financial data needed to make accurate predictions. While one accountant found the available data sufficient for their assessments, others highlighted missing elements that could have improved their evaluations. Specifically, they pointed out the absence of:

- Reasons for existence of the company
- Added value
- Shareholders
- Financial data over multiple years (past, present, and future)
- Bank covenants
- Liquidity forecasts
- Management's financial expectations and future plans
- Full balance sheet and P&L statements

Additionally, some participants indicated that errors or inconsistencies in the dataset led to doubts about the reliability of certain financial data. However, these errors were present in the raw data rather than in the analysis or the model itself. While this might have affected user perception, it aligns with the challenges auditors face when working with incomplete or inconsistent financial statements in real-world situations and should not impact the results of the case study.

4.3.4 Satisfaction with the model

As part of the evaluation, the LLM group was also asked about their experience and satisfaction with the model. Appendix C illustrates the answers of the accountants. The results show that all participants incorporated the model into their predictions, although four out of five did so only to a limited extent. Their primary reason was that, while the model provided an initial trend and insights, they still preferred to conduct a deeper analysis beyond what the model summarized. One accountant explicitly mentioned that they attempted to rely on it as little as possible.

One of the most valuable aspects of the model, according to the LLM group, was its ability to highlight key financial indicators and ratios. Additionally, two accountants found the textual summary particularly useful, especially for capturing insights not directly reflected in the ratios or balance sheet.

Regarding potential improvements, the accountants provided several suggestions. One recommended removing the auditor's report and focusing solely on the balance sheet and P&L statements, as the model sometimes considered macro management risk factors mentioned in the report that were not always relevant. Another suggestion was to structure the risk assessment into clear categories (e.g., low, medium, high) before presenting further details. The remaining three accountants had no further feedback on the model.

When asked whether they would use the model's predictions in real-world assessments, all participants answered 'yes'. They saw it as a valuable early warning tool that could complement their own analysis, as long as it could be integrated into their workflow with minimal effort.

Chapter 5

Discussion

In this chapter, we discuss the previously presented results and connect them to related work. Furthermore, we reject or accept the hypotheses and highlight limitations of this research. Additionally, a section is appointed to discuss the use of the DSPy framework. Lastly, some recommendations are given to further refine the model and potentially use it within the host company.

5.1 Model accuracy

5.1.1 Feature selection

The results showed the differences in model performance with regards to feature selection and the various optimizers. First of all, the selection of the features impacted the accuracy and false negatives of the model to a relatively large extent. Taking all features into account when predicting bankruptcy resulted in a 5% lower accuracy than taking a subset, which leads us to believe that adding too much information to the model weakens its performance. This was also found by Bellovary et al. [12], who concluded that, for bankruptcy prediction, "higher model accuracy is not guaranteed with a greater number of factors." Their analysis showed that some models with two factors were capable of predicting bankruptcy as accurately as models with 21 factors [12]. Moreover, the baseline with mutual selection-based features scored 76% accuracy, whereas the features chosen by accountants achieved 78% accuracy. This underlines the idea that while LLMs can generalize representations without explicit training [16], their effectiveness in bankruptcy prediction improves significantly with expert-driven feature selection, highlighting the role of domain knowledge in refining model performance.

A closer examination of the feature subsets, which are presented in Appendix A, reveals a potential reason for this difference. The mutual information score selection includes features such as *'ZKH: Psychiaters fte PUK en PAAZ (loondienst inhuur vrij beroep)'* and *'STZ JN'*, which contain a large number of NaN values. These missing values were replaced by the median or 'Unknown' for quantitative and qualitative features, respectively. Their inclusion in the selected subset could have been caused by data imputation rather than their predictive value. However, many of the remaining features closely align with those chosen by accountants, indicating that mutual information-based selection is relatively effective, although not as refined as expert-based feature selection.

5.1.2 Optimizers

With regards to model accuracy, the key insight is that, regardless of which optimizer is used, the model performs relatively well: within a range of 76% - 79%. As said before, the differences in performance across optimizers were minimal. This suggests that while optimization techniques can refine model performance, their overall impact remains limited within this research. This goes against the expectations of the study by Eulerich et al. [28], who showed that enhancing the ChatGPT model resulted in significantly better performance. However, Sarmah et al. [61] found a similar effect of the DSPy optimizers, which only managed to improve the accuracy from 82.13% to 85.87% with the best performing optimizer. A possible explanation is that Eulerich et al. [28] employed different optimization techniques, which are not directly comparable to the DSPy optimizers used in our study and the one by Sarmah et al. [61].

Eulerich et al. [28] incorporated ChatGPT 3.5 and 4, applied both zero-shot and 10-shot learning, and explored the use of ReAct and non-ReAct prompting. In contrast, the DSPy optimizers focus on few-shot learning, prompt optimization, or a combination of both. Since Eulerich et al. [28] used various model versions and reasoning techniques alongside prompt engineering, their optimizations may have had a broader impact on performance. This could explain why their study observed stronger improvements, whereas the optimization techniques in our research and the study by Sarmah et al. [61] led to only minor accuracy gains.

Despite the marginal differences in accuracy, it is important to see the differences between our optimizers in terms of optimization approach. With the exception of COPRO, all optimizers use some form of few-shot learning. MIPROv2 goes a step further by not only incorporating few-shot learning but also optimizing the prompt provided to the model. COPRO, on the other hand, optimizes the prompt without using few-shot learning. Notably, despite the absence of example-based learning, COPRO outperforms all other optimizers, highlighting the effectiveness of prompt optimization alone. However, since the LabeledFewShot optimizer has a similar accuracy and just a slightly higher number of false negatives, it cannot be concluded that optimizing the prompt alone, as COPRO does, is the best approach. Instead, the results suggest that prompt optimization is just as critical as providing examples to the model, and the optimal strategy may lie in balancing both elements effectively. Interestingly, MIPROv2 attempts to combine both few-shot learning and prompt optimization, yet performs worse than both COPRO and LabeledFewShot. The similar study by Sarmah et al. [61] showed that, in fact, MIPROv2 scores the best with an accuracy of 85.87% when tested on various datasets. COPRO is a low performer, with an accuracy equal to the DSPy baseline of 82.13%. They argued that this was due to the variable structure of the data [61]. This raises the possibility that our hyperparameter optimization within MIPROv2 requires further refinement. As limited research exists on the evaluation of the DSPy optimizers, further investigation is needed to fine-tune their performance and determine the optimal balance between prompt optimization and few-shot learning. At this stage, it remains inconclusive which optimization approach is most effective for bankruptcy prediction.

5.1.3 COPRO optimizer

The previous chapter outlined the prompts generated and tested by the COPRO optimizer. Given that the model learns from each iteration, one might expect to see a continuous improvement in accuracy scores of prompts over the iterations. This trend is evident for folds 1 and 2, where accuracy scores generally increase with each iteration. However, fold 3 deviates from this pattern, as the scores slightly decline across iterations. One possible explanation for this could be that the initial prompts generated in fold 3 were already fitting to the task at hand, meaning that further modifications did not lead to improvements. As a result, the optimizer may have struggled to refine the prompts in meaningful ways beyond the initial generation. This could be deduced from the fact that the first iteration contained two prompts with higher accuracies than the best prompt from the first fold.

Another interesting observation is that the COPRO optimizer generated identical prompts in different folds. Since there is no memory between folds or between API calls, this suggests that the model follows a certain consistent pattern or strategy when optimizing prompts. This finding could imply that the optimizer has identified a few key features of prompt formulation that are particularly effective, which it replicates across different training iterations.

Furthermore, the relationship between prompt length and accuracy is not as straightforward as one might expect. For example, prompt 3 from fold 1, which was relatively short, achieved an accuracy of 77.8%, while prompt 7 from the same fold was notably longer and only scored 74.1%. This suggests that factors other than prompt length, such as clarity or wording, play a more significant role in determining the effectiveness of the prompt. Therefore, it is important to consider the content and structure of prompts rather than only their length when evaluating their performance.

Examining the content of the prompts, it can be seen that the best-performing prompts from fold 1 and 2 both focus on the generation of the PNG summary. Since this summary is already given to the model as data, this should not impact the evaluation and should not even be the task of the model. However, seeing that they perform well in comparison to prompts that do not specifically focus on the creation of the summary, it might have an influence after all. This indicates that the PNG summary may unintentionally influence the model's performance, highlighting the need for further research on its impact on the evaluation process.

Lastly, considering that the scores range from 72.2% to 79.6%, we can conclude that the structure and content of the prompt have a moderate impact on the accuracy of bankruptcy prediction. This is consistent with the overall findings from the optimizers, where differences in accuracy were not particularly large either (76% - 79%). Therefore, while fine-tuning prompts with the COPRO optimizer shows some impact, it does not appear to drastically improve prediction accuracy.

5.2 Boundaries of the model

The results of the temporal analysis indicate that the model's accuracy declines as the prediction time frame increases, suggesting that the model becomes less reliable for long-term bankruptcy prediction. This aligns with expectations, as financial conditions and external factors (e.g., policy changes or market conditions) can shift significantly over time. These findings are consistent with prior research on bankruptcy prediction, which typically focused on one to two years before bankruptcy [12]. However, Bellovary et al. [12] also highlight that some studies have successfully predicted bankruptcy three to five years in advance. The inability of this model to match the five-year time frame of some prior studies may be due to our dataset constraints or model design choices. Specifically, the accuracy dropping below 50% at five years before bankruptcy indicates that the model is no longer extracting meaningful predictive patterns and performs worse than random guessing. This suggests that financial statement data alone may not provide sufficient predictive power for long-term forecasting. Alternative approaches, such as integrating external economic indicators, industry-specific risk factors, or qualitative auditor assessments, may be necessary to improve long-term predictions. However, despite the observed decline in accuracy over time, the model remains useful for predictions up to three years before bankruptcy with an accuracy of 85%, making it a valuable tool for early risk assessments in auditing. Additionally, the observed decline in performance with the increased time frame also shows that the model is learning from meaningful patterns in the data, rather than producing random outputs. This strengthens the confidence in the model's performance.

It is important to interpret these findings with caution due to the dataset limitations. The sample size of 20 bankrupt organizations per year may be insufficient to ensure the reliability and robustness of the model's long-term predictions. A larger dataset would be necessary to confirm whether the observed trend holds across a larger population.

5.3 Case study

5.3.1 Difference between model and accountants

Overall, the model's predictions closely align with those of the accountants. The organizations that the model struggled to predict accurately were also the ones that were challenging for the accountants. This suggests that the model considers the right indicators from an auditing perspective, based on the financial data provided, and selects similar cutoff points for determining bankruptcy as accountants. Notably, there are instances where the model even outperformed the accountants. For example, organization 15 from the case study was predicted by all accountants to avoid bankruptcy, while the model correctly predicted that it would go bankrupt. To further evaluate this distinction, the financial data of the organization is visualized in Figure 5.1. Additionally, the predictions of both the model and the accountants for this organization are presented in Table 5.1 and 5.2, respectively.

The reasoning shown in Table 5.1 is the thought-process the model used to get to the outcome of the prediction, while the explanation is where the model tries to explain the answer it gave. The former is a part of the Chain of Thought prediction module of the DSPy framework, whereas the latter is an output field we included to ensure that the model could explain its predictions. The reasoning part of the model shows that it focused mainly on the negative operating and net result, as well as the solvency rate and decline in equity. The liquidity ratio is mentioned as well, but according to the model, the positive number does not weigh up against the other negative figures.

The accountants focus mostly on the fact that there is no (long-term) debts and that the liquidity ratio is high, suggesting that equity is sufficient enough to cover for short-term debts and specifically the disappearance of provisions.

When comparing the model's and accountants' explanations, it becomes evident that the model provides a more comprehensive assessment of the company's financial health. While accountants focused on short-term positive factors, such as liquidity and reserves, they underestimated long-term risks, including sustained losses, declining equity, and a concerning solvency ratio. In contrast, the model accounted for these critical negative indicators, accurately predicting the company's potential bankruptcy. This suggests that a broader analysis of financial health is crucial for identifying long-term risks. The model's ability to capture these risks likely stems from its transformer architecture, which leverages self-attention to process complex and multi-layered financial data as a whole [29]. By simultaneously analyzing relationships between various financial indicators, the model can detect patterns and dependencies that might be overlooked in a more fragmented approach. Furthermore, this comparison shows the challenges humans face in processing larger volumes of unstructured data in the case study. In contrast, LLMs excel in systematically analyzing such information [9], suggesting their potential value in enhancing the efficiency and reliability of going concern assessments by focusing on the information processing part of the assessment.

PASSIVA

	31 december 2021		31 december 2020	
	€	€	€	€
Eigen vermogen				
geplaatst kapitaal	18.000		18.000	
overige reserves	1.143.003		1.442.907	
		1.161.003		1.460.907
Voorzieningen				
overige voorzieningen		-		187.983
Kortlopende schulden				
crediteuren	1.313		4.013	
overige schulden en overlopende passiva	9.075		16.860	
		10.388		20.873
		1.171.391		1.669.763

Financiële data

Bedrijfsresultaat: -304033.0
Langlopende schulden: 0.0
Liquide middelen: 138758.0
Liquiditeitsratio: 112.8
NVTZ Omzetklasse: € 0.7 tot € 10 mln.
Overige schulden: 0.0
Resultaat: -299904.0
Resultaatratio: -0.266
Schulden aan banken: 0.0
Schulden aan groepsmaatschappijen: 0.0
Schulden aan leveranciers en handelskredieten: 1313.0
Schulden aan participanten en aan maatschappijen waarin wordt deelgenomen: 0.0
Schulden ter zake pensioenen: 0.0
Schulden uit hoofde van financieringstekort: 0.0
Solvabiliteit (debt ratio): 1.028
Som der bedrijfslasten: 1433240.0
Som der bedrijfsopbrengsten: 1129207.0
Totaal activa: 1171391.0
Totaal passiva: 1171391.0
Year: 2021

Figure 5.1: Financial data of organization 15 from the case study.

Table 5.1: Model's prediction and explanation for organization 15 (Actual: Yes).

Category	Output
Predicted Outcome	Yes
Reasoning	The company has a negative operating result of -€304,033 and a negative net result of -€299,904, indicating sustained losses. The liquidity ratio of 112.8 is high, suggesting the company can meet its short-term obligations. However, the solvency (debt ratio) of 1.028 is concerning, as it indicates that liabilities exceed equity. The decline in equity and the disappearance of provisions point to a deterioration in financial health. While total liabilities have decreased, this is mainly due to a reduction in equity and provisions, which is not a positive sign.
Explanation	The company is experiencing sustained losses, with a negative operating result of -€304,033 and a negative net result of -€299,904. The solvency (debt ratio) of 1.028 means that liabilities exceed equity. The decline in equity and the disappearance of provisions suggest a worsening financial condition. Despite a high liquidity ratio, these factors indicate an increased risk of bankruptcy within two years.

Table 5.2: Accountants' predictions and explanations for organization 15 (Actual: Yes).

Category	Output
Predicted outcome	No
Explanations control group	"I assume the loss in 2021 is incidental, as there are no debts owed to banks or other financiers. Management can still reduce costs and secure financing to sustain operations for at least another two years."
	"The financial result is poor, but liquidity is still sufficient. The disappearance of provisions is not necessarily a sign of a deteriorating financial position."
	"No long-term debt. Short-term liabilities can be covered using liquid assets. However, operating results are declining."
	"There is an issue with provisions, but there are still significant reserves. No reason to expect bankruptcy within two years."
	"Equity is so high relative to debt that I do not anticipate bankruptcy. Liquid assets are sufficient to repay debts in full. I do not see which party would file for bankruptcy."
Explanations LLM group	"Very strong equity and very limited debt."
	"Negative result, but healthy equity, no long-term debt/financing credits."
	"Very limited debt in relation to liquidity. Equity sufficiently covers negative results. Very strong/high solvency."
	"Strong equity, almost no debt. Enough capital to absorb possible losses in the coming years."
	Did not evaluate

This also raises the question of whether accountants may be overly optimistic compared to the model when assessing an organization's financial health. The confusion matrices below (Table 5.3 and 5.4) suggest that this is indeed the case. On average, accountants predicted bankruptcy only 36% of the time, while the model did so 55% of the time. This indicates that the model classifies organizations as bankrupt more frequently, whereas accountants may be more reluctant to make such a prediction. One possible explanation is that, in practice, bankruptcies are relatively rare, whereas bankrupt and non-bankrupt organizations were evenly distributed in the sample of the case study. As a result, accountants may be less inclined to predict bankruptcy as frequently based on their previous experiences. In addition, accountants often lack a structured feedback loop regarding the accuracy of their predictions. Unless they assess the same organization over multiple years, they may not always learn whether their previous assessments were correct, which could reinforce a conservative approach in their predictions. The LLM, on the other hand, is pre-trained on large amounts of general data, allowing it to identify patterns that might not be as apparent to accountants.

Table 5.3: Confusion matrix of model.

Actual \ Prediction	No	Yes
No	40%	10%
Yes	5%	45%

Table 5.4: Confusion matrix of accountants.

Actual \ Prediction	No	Yes
No	46%	4%
Yes	18%	32%

Another interesting case is organization 1, where only one accountant of the LLM group predicted incorrectly. They mentioned the positive equity of the organization as a contributing factor to non-bankruptcy. However, the financial data shows that the equity is negative, which is also recognized by the model. This could suggest an error in reading, leading to a false prediction. While such errors are inherent to human evaluators, they are unlikely to occur in the model's analysis, highlighting a potential advantage of LLM assessments in bankruptcy prediction.

Our findings substantiate the study by Kim et al. [37], who found that chatGPT-4 was able to predict future earnings more accurately than human analysts, especially in complex scenarios with which analysts struggled. Although our model was not necessarily able to predict better than the accountants overall, it still matched the performance of the best and performed better than the others, thus supporting the findings of Kim et al. [37]. Additionally, Kim et al. [37] emphasized that LLMs can be particularly valuable in assisting human analysts when they are underperforming due to bias or disagreement, offering a more objective and consistent perspective. Our findings are in line with this idea as well, as the model's predictions provided valuable insights in situations where human analysts appeared hesitant to predict bankruptcy. This suggests that LLMs could serve as a supportive tool for human judgment and decision-making, particularly in scenarios where human bias or hesitation might lead to inconsistent assessments.

5.3.2 Difference between control and LLM group

The results of the case study revealed no significant difference in performance between the control group and the LLM group. Both groups demonstrated equal accuracy and speed in predicting bankruptcy. This suggests that incorporating the model's predictions and the summary of the auditor's report did not improve the bankruptcy predictions for the LLM group. There are several reasons that could contribute to this lack of improvement. First of all, the accountants were unfamiliar with the model's predictions and had to learn how to integrate them into their evaluations during the case study, impacting the speed of their assessment. Furthermore, the accountants incorrectly deviated from the model several times. However, after working with the model for a longer period of time, they might learn to trust it more, potentially leading to better outcomes in future bankruptcy predictions. This is underlined by Anica-Popa et al. [5], who found that accountants need certain skills and knowledge to effectively use generative AI within auditing. However, it is worth noting that the LLM group did not perform worse than the control group, meaning the model did not negatively impact the accountants' predictions. If the assumption holds that the accountants will be more accurate and faster with more knowledge on the model, the model has the potential to have a positive impact on the accountants' evaluations.

The results for the edge cases, where accountants disagreed on a prediction, suggest that the LLM group has a slightly higher prediction accuracy than the control group. This highlights a few notable observations. One example is organization number 3, where the model correctly predicted bankruptcy but two accountants of the LLM group deviated from this prediction, as well as four of the five accountants in the control group. Table 5.5 shows the explanations that the accountants gave for their predictions. They mostly rely on the positive liquidity ratio for the argumentation, but also mention aspects like personnel costs, solvency ratio, and impairments. Interestingly, this is mentioned in both the explanations for bankruptcy as for non-bankruptcy. Moreover, the LLM group demonstrates a more critical approach to certain metrics, with some accountants explicitly questioning the validity of the liquidity ratio and pointing out inconsistencies in the financial data. In contrast, the control group appears to place greater trust in the liquidity ratio as a decisive factor in their predictions, often using it as a justification for non-bankruptcy despite other warning signs such as negative results and high personnel costs. Another noticeable trend is that the LLM group more frequently considers long-term sustainability, discussing the impact of high personnel costs and external hiring on future equity levels. Meanwhile, control group accountants focus more on immediate liquidity and short-term obligations, sometimes explicitly stating that their assessment does not extend beyond a one-year horizon. This suggests that the LLM group's exposure to the model may have positively influenced them to take a broader view of financial risk, whereas the control group remains more aligned with traditional auditing time frames and criteria. As mentioned in Section 5.3.1, this broader view of financial risk could be a benefiting factor to the accuracy in bankruptcy predictions. This is also what the model incorporates in its predictions.

However, from this singular example it cannot be concluded that the model positively influenced the LLM group in their decisions. For instance, for organization 1, one LLM accountant predicted that the organization would not go bankrupt based on positive equity, despite the model indicating negative equity. This accountant overlooked this key factor, which suggests that while the model may help guide decisions, it does not guarantee that accountants will always incorporate its insights correctly. Nevertheless, these findings highlight a valuable opportunity for accountants to learn how to integrate the model into their workflows to be able to apply it more effectively.

Table 5.5: Predictions and explanations from accountants for organization 3 (Actual: Yes).

Group	Prediction	Explanation
LLM	No	The liquidity ratio is positive, so there is no immediate liquidity problem. I consider it likely that management can still take measures to reduce costs.
	No	No indications of a decline in business revenues or other significant cost increases. Fixed assets are limited, indicating a labor-dependent organization. This implies personnel costs. An increase in personnel costs likely suggests business expansion. There are no signs that revenues will decline due to market conditions or price reductions.
	Yes	Low cash reserves, poor results relative to revenues. I also do not understand the liquidity ratio—why is it 1.56?
	Yes	It could be possible. Creditors are quite high, and personnel costs are also high. Therefore, equity will decline rapidly.
	Yes	In addition to an increase in full-time employees, there are also higher costs for non-payroll personnel, which are apparently not sufficiently compensated by revenues. The use of non-payroll personnel also indicates a shortage of internal staff, possibly due to illness replacements. Given the low cash reserves, the company cannot continue operating like this for another two years.
Control	No	The result is negative, and the profitability ratio is low. However, they have a liquidity ratio of 1.56. If this is the current ratio, then they have sufficient short-term receivables to pay off their debts. So, they should still be able to settle their obligations. The solvency ratio is also low, indicating they are not in a dire financial situation. However, personnel costs are high, which could lead to problems, but not within the next two years.
	No	Based on the liquidity ratio, the company can meet its obligations. As accountants, we do not assess a two-year horizon, only one year after the audit opinion. Additionally, there is insufficient information for a proper assessment.
	No	The company is operating at a loss, but there are no impairments, which could have been a warning sign (e.g., impairment due to declining activities).
	No	Liquidity ratio is well above 1, indicating that receivables will be collected in the short term. Non-payroll personnel can be let go. Recovery is possible. Too premature to conclude bankruptcy despite the creditors' position.
	Yes	Negative result, and most of the expenses consist of personnel costs. The creditors' position is quite high compared to the available cash.

Regarding the ease of the assessment, the LLM group found it slightly easier than the control group. As previously mentioned, many accountants described the lack of certain financial data as a key reason for the difficulty in making their assessments. This was particularly noted by accountants in the control group who found the task somewhat difficult. While the LLM group also acknowledged this issue, they still considered the assessment relatively easy. This implies that the additional information provided by the model's predictions may have been perceived as helpful, potentially simplifying the evaluation process. Accountants in the LLM group specifically noted how the model's ability to highlight key financial ratios and offer concise summaries of the organization's financial health was particularly useful. This extra guidance might have helped them focus on the most critical indicators for predicting bankruptcy, reducing the cognitive load of the assessments.

5.3.3 Satisfaction with model

This section is based on the questions in the evaluation form regarding the model, which were only answered by the LLM group.

The fact that most accountants incorporated the model only to a limited extent suggests that while they found it useful, they still prioritized their own expertise and judgment. The reluctance of one accountant to rely on it may also indicate concerns about over-reliance on AI or a preference for traditional auditing methods. This aligns with existing research on AI adoption in auditing, where professionals often have a skeptical view of the use of AI [49].

The suggestion to remove the auditor's report and focus only on financial statements indicates that some model inputs may introduce noise instead of valuable information. This aligns with the findings mentioned in Section 5.1, that adding more data does not necessarily improve a model, as it performed better with a selected subset of features rather than the whole set. Additionally, a participant suggested that structuring risk assessments into clear categories could enhance interpretability, aligning the model more closely with how auditors typically assess risk. This highlights the importance of presenting AI-generated insights in a structured and intuitive way, as well as ensuring compatibility with existing auditing workflows.

Despite limited reliance on the model during the evaluation, all participants indicated that they would use the model in real-world assessments, emphasizing its role as an early warning tool. This suggests that LLM-based models can be effectively integrated into auditing workflows, provided they require minimal effort to use. It does go against the statement that accountants are skeptical towards the adoption of AI in first paragraph. However, it could be explained by the fact that the case study was on a voluntary basis, introducing a bias towards accountants that have some sort of familiarity or affinity with AI. This should be taken into account when interpreting the results, as broader adoption within the auditing field may require additional training to improve accountants' confidence in the reliability of the model.

The most appreciated features of the model were the highlighted key financial indicators and provided textual summaries. The textual summary, in particular, seems to add to the traditional quantitative analysis by capturing qualitative insights that might otherwise be overlooked. This suggests that the main advantage of the model lies in enhancing information processing rather than replacing expert judgment. This was also evident in the comparison between the model and the accountants, where accountants struggled to integrate all available information, whereas the model provided a broader and more comprehensive assessment of the organizations' financial health.

5.4 Hypotheses

The research question and the underlying hypotheses were introduced in Chapter 1. This section goes into the acceptance or rejection of the hypotheses to be able to answer the research question later in the conclusion.

H.1 Programmatic prompt-tuned LLMs are more effective than traditional auditing methods in assessing the going concern.

The results from the case study showed that the LLM performed similarly to the highest-performing accountants in both the LLM and control groups when predicting bankruptcy. Additionally, the LLM group demonstrated no significant advantage over the control group. This suggests that traditional auditing methods are just as effective regarding the prediction of bankruptcy in the healthcare sector. Consequently, we reject this hypothesis, as the model did not outperform the traditional auditing methods in going concern assessments.

H.2 Programmatic prompt-tuned LLMs can shorten the current process of assessing the going concern while maintaining accuracy.

The case study findings showed that the LLM group required the same amount of time to evaluate bankruptcy as the control group. Although the LLM demonstrated accuracy in the predictions, it did not reduce the time spent by accountants during the evaluation. Since the process of the going concern assessment was not shortened, the second hypothesis is rejected as well.

H.3 Accountants perceive LLMs as a valuable tool to assist in going concern assessments.

Feedback from the case study evaluation revealed a positive perception among accountants regarding the LLM. Participants expressed willingness to integrate the model into their workflows, acknowledging its potential as a supportive tool in going concern assessments. However, it is important to note that these responses might be influenced by a sample bias towards accountants with a higher affinity for AI tools. Therefore, further validation a larger sample of accountants with different AI affinity levels would be beneficial to confirm these findings. Nevertheless, within the context of this research, the accountants perceived the LLM as a valuable tool, which is why we accept this hypothesis.

5.5 DSPy framework

Beyond the broader discussion on the practical use of LLMs, it is valuable to examine the DSPy framework and its advantages and limitations. Chapter 3 outlines our rationale for selecting DSPy to optimize the LLM, but there are key considerations when implementing this framework.

One of the primary challenges is its steep learning curve. Setting up DSPy, even for a small prediction model, requires significant effort. However, its strength lies in the second phase. Once the framework is in place, modifying components, such as prediction modules or optimizers, becomes more straightforward. This modularity aligns with the creators' intent to make it easier to refine and adapt models over time [36]. However, the decision to use DSPy should be based on the scale of the project. For a simple prediction model, writing custom code may be sufficient, whereas larger and more complex models can benefit from DSPy's structured approach.

Additionally, the DSPy framework is continuously updated and refined. This ongoing development is both an advantage and a hindrance. On the one hand, frequent updates improve the framework over time, enhancing its capabilities. On the other hand, these updates sometimes introduce modifications that require existing code to be adjusted to maintain compatibility.

In comparison, another framework for handling LLMs is Langchain. Langchain and DSPy both serve as frameworks for utilizing LLMs, but they differ in their approach and core functionality. Langchain is designed as a flexible, general-purpose framework that enables seamless chaining of multiple components, making it well-suited for the integration of various tools and external models into their applications. Its strength lies in its adaptability and ability to facilitate complex workflows. In contrast, DSPy takes a more structured, programmatic approach, focusing on modular and declarative programming techniques for instructing LLMs. It integrates prompting, fine-tuning, reasoning, and retrieval augmentation within one framework, offering a more automated and Pythonic method for optimizing model performance. While Langchain prioritizes flexibility and broad usability, DSPy emphasizes structured optimization and automation [31].

Additionally, Schnabel and Neville [62] introduced an alternative framework called SAMMO, which follows a similar direction. SAMMO uses symbolic prompt programs to represent valid prompt structures, enabling flexible transformations and efficient search for optimized prompts. In their comparison with DSPy's COPRO, they found that DSPy underperformed because its prompting often caused the model to deviate from the expected output format [62]. This suggests that the SAMMO framework is a promising alternative to the DSPy framework.

Overall, while DSPy presents challenges in the initial setup and requires ongoing maintenance due to frequent updates, its modular and automated optimization capabilities make it a powerful tool for refining LLMs. Compared to more flexible frameworks like Langchain, DSPy provides a structured approach that is well-suited for systematic model optimization, supporting our choice to use it in this study. However, given SAMMO's promising approach to symbolic prompt program search, future research should explore whether it offers even better optimization strategies for improving LLM performance.

5.6 Limitations

Despite the contributions to the literature on practical use of LLMs within auditing, there are various threats to the validity of this research that will be discussed in this limitation section.

A key limitation lies in the dataset itself. The financial data of healthcare organizations from 2018 to 2023, during the COVID-19 pandemic, may not accurately reflect long-term trends, as the pandemic introduced exceptional financial pressures. Additionally, the dataset contains a significant number of missing values that required pre-processing, which may have influenced the final outcomes. Another constraint is the limited number of bankrupt organizations within this time frame. Since relatively few Dutch healthcare organizations have gone bankrupt in recent years, only a small subset of the dataset could be used, making it difficult to assess the model's generalizability. Furthermore, most bankrupt organizations in the dataset had annual revenues below 10 million euros, meaning the findings may not be fully representative of larger healthcare organizations. Moreover, since healthcare organizations are typically non-profit [69] and government-supported [72], financial distress may not lead to bankruptcy as quickly as in profit-driven sectors.

As a result, the findings regarding the performance of the LLM and its optimizers may not be directly applicable to other industries. However, some aspects of this study remain generalizable. For instance, the importance of structured and unstructured data integration, the model's ability to avoid human errors in data processing, and its comprehensive overview of financial health are relevant across different sectors. Furthermore, the qualitative insights from accountants, such as their perceptions of LLM-generated summaries and their willingness to incorporate the LLM into their workflow, are not healthcare-specific. These broader implications suggest that, while the specific predictive performance of the model may vary across industries, its ability to enhance financial assessments and decision-making processes holds potential beyond the healthcare sector.

There also exist some limitations regarding the integrity of the dataset. The financial data used in this study was originally acquired by BDO from an external provider, limiting our ability to verify its accuracy. Although much of this information is publicly available, manually cross-checking all data points was not feasible within the scope of this research. The accountants involved in this study also pointed out inconsistencies, such as solvency ratios that did not align with reported private equity and total assets. While these inconsistencies did not necessarily affect the internal validity of the study—since both the LLM and accountants worked with the same dataset—they do impact the broader conclusions that could be drawn about the financial health of the organizations.

Beyond data-related challenges, there are methodological considerations that could have influenced the results. Given the high computational costs of the model's optimizers, the number of features used in the model had to be restricted. While this was a necessary step, the selection of features remained somewhat subjective and could potentially affect the model's performance.

Additionally, there is a small chance that an accountant recognized one of the organizations in the case study based on the financial data provided. If this were the case, prior knowledge could have influenced their judgment. However, since no indication of this was found in the qualitative explanations, we assume that this was not a significant issue.

The construction of the control and experimental groups presents another limitation. While efforts were made to ensure a comparable level of expertise, the control group included two accountants who did not hold an RA (Registered Accountant) title, whereas the LLM-assisted group had only one accountant without this title. Although all participants had relevant experience, this difference in professional qualifications could have had a minor impact on the results.

Furthermore, a larger group of accountants was initially invited to be involved in the case study, but only those who were willing ultimately participated. This willingness could indicate that participants had a greater interest in or affinity with AI, potentially introducing a bias. Those more open to AI-driven tools may have evaluated the model more favorably, while more skeptical accountants may have opted out, affecting the generalizability of the findings.

Finally, while LLMs demonstrate impressive capabilities, they also have several limitations that must be acknowledged. These include the risk of biased outputs due to potentially biased training data and a tendency to 'hallucinate' or generate incorrect information [32]. Moreover, there are challenges in interpreting how the model arrives at specific predictions. This black-box nature of LLMs makes them difficult to explain [32], which may reduce trust in sensitive contexts such as auditing. This underscores the importance of human oversight and careful integration into existing auditing processes.

These limitations should be considered when interpreting the results and drawing conclusions about the applicability of LLMs in going concern assessments. However, this research represents a small-scale experiment meant to provide initial insights rather than definitive conclusions. The findings serve as a baseline that future studies can build upon and validate on a larger scale.

5.7 Recommendations

As this research was conducted in collaboration with BDO Netherlands, we provide some recommendations for the use of LLMs within their auditing processes. These suggestions can be categorized into model design improvements and recommendations for the integration into existing workflows. They include both adjustments that were outside the scope of this research as well as insights derived from the case study that could not be implemented within the current time frame.

5.7.1 Model design recommendations

With regards to the design of the model, the first recommendation is to ensure the validity of the data used in the model to guarantee its accuracy and reliability. In the case study, it was noted several times that the data did not always align, though this did not affect the comparison between the model and accountants. However, given the critical role financial data plays in decision-making, it is essential to ensure that the input data is robust and trustworthy when using the LLM in a real-world scenario.

In terms of the model's output, instead of requesting a binary 'Yes' or 'No' bankruptcy status, it would be more beneficial to ask for a probability or risk category. This would provide auditors with more nuanced insights and allow them to assess risks with greater flexibility and precision. This approach aligns with feedback from accountants during the case study, who indicated that more detailed risk assessments would enhance their decision-making.

Another key improvement is to minimize false negatives in the model's predictions. The model could be adjusted to reduce the likelihood of overlooking potential risks, which could have serious consequences. While maintaining overall accuracy, emphasizing this goal in the prompt would ensure that the model aligns with the host company's priorities. The initial prompt used in this study did not specifically address this concern, but its inclusion could enhance the relevance of the model's results.

Additionally, incorporating the entire auditor's report into the model could be valuable. Given that accountants found the textual summaries beneficial for their assessments, using the full report might provide a more comprehensive context for the model's predictions, potentially increasing the model's accuracy and usefulness. However, this should be experimented with, as the results indicated that including all features reduced accuracy. Similarly, integrating the full report may also introduce noise, potentially decreasing performance rather than improving it.

Another valuable addition would be integrating financial data from multiple years, allowing for a better understanding of an organization's financial progression. This was particularly noted in the case study, where historical trends were missing from the assessment.

Finally, BDO could focus on using the best prompt from COPRO for the model in other optimizers, as this has shown positive results in generating accurate and insightful responses. By experimenting with other optimizers, BDO could determine the configuration that leads to the highest performance and fits the specific needs of their auditing process.

5.7.2 Model integration recommendations

When integrating LLMs into BDO's auditing processes, there are several considerations that can streamline the adoption. First of all, for healthcare organizations, which typically have publicly available financial data, using a LLM for audits is relatively straightforward. However, for dealing with confidential data, BDO should consider utilizing their own internal LLM, a customized version of the ChatGPT model. This would require some modifications to ensure data privacy and compliance, but it would provide a more secure way to use LLMs without compromising confidentiality.

In addition, all accountants from the case study were positive about using the LLM as part of their audits and they particularly valued its potential as an early warning tool. The model could provide an alert to accountants about potential financial risks early in the assessment process, acting as a baseline for further review. This would allow auditors to focus on areas that require closer review, streamlining their workflow and improving decision-making. The accountants also highlighted the need for the LLM to be easy to use with minimal effort. Therefore, the tool should be designed to enhance the audit process, rather than complicate it.

To successfully integrate the LLM into BDO's auditing processes, it is essential to align the adoption with their existing innovation process. BDO follows a structured approach to innovation, starting with a small-scale pilot to gather feedback and refine the tool. Depending on the extent of necessary adjustments, this is followed by either another pilot or a broader experiment within a specific sector or audit process. If successful, the tool is tested in practice by accountants, with potential oversight from regulatory entities like the AFM and internal quality control to ensure reliability and compliance. This is particularly important with machine learning and LLM-based tools, since they sometimes lack the explainability accountants depend on. Given this process, our case study can be viewed as the initial pilot. A logical next step would be to conduct a larger-scale experiment within a selected audit domain, ensuring alignment with BDO's internal standards and regulatory expectations. This would provide deeper insights into the tool's effectiveness and facilitate further refinements before broader implementation.

Lastly, to ensure successful integration, accountants should receive training on using the LLM effectively. Anica-Popa et al. [5] found that accountants require specific skills and knowledge to effectively utilize generative AI in auditing. This is also reflected in the case study, where the accountants using the LLM were just as accurate and fast as the control group, despite having no prior experience with the model. With proper training and more experience, they could potentially become even more effective and efficient, leveraging the model's strengths while maintaining professional skepticism. Providing training sessions or guidelines will help accountants to critically assess the model's outputs rather than rely on them blindly, both ensuring accuracy as well as adhering to the auditing standards.

By incorporating these recommendations, BDO can design a model that not only meets the needs of accountants but also adds significant value to the auditing process. With the use of validated data, clearer risk categorization, and the integration of qualitative insights, the LLM-based tool can enhance decision-making by making audits more proactive and informed.

Chapter 6

Conclusions

6.1 Summary

The COVID-19 pandemic caused financial instability in the healthcare sector, underscoring the importance of robust going concern assessments [71]. While traditional machine learning approaches struggle with analyzing unstructured financial data in such going concern assessments [43], LLMs have shown potential in addressing these complexities [9]. However, their application in auditing, particularly for assessing the financial continuity of healthcare organizations, has remained largely unexplored [38] [72].

This study investigated the feasibility of using LLMs to predict bankruptcy in healthcare organizations by combining structured financial data with unstructured auditor's reports. Following a design science approach and using the DSPy framework, we developed a LLM-based model designed to predict bankruptcy. Its performance was evaluated in a case study, comparing its predictions with those of accountants and assessing its impact on assessment accuracy and efficiency when used by accountants.

The findings showed that feature selection proved crucial, with expert-selected features leading to higher model accuracy than mutual information-based selection, highlighting the value of domain expertise. While different optimizers showed minimal impact on overall performance, COPRO outperformed others with a 79% accuracy and low false negatives, suggesting that effective prompt optimization alone can enhance predictive accuracy. However, the results indicate that an optimal bankruptcy prediction strategy may require a balance between prompt optimization and few-shot learning, as the LabeledFewShot optimizer also achieved an accuracy of 79%.

When examining the boundaries of the model, it performed well for short-term bankruptcy predictions, maintaining 85% accuracy up to three years before bankruptcy. However, its reliability declined over longer time horizons, becoming less accurate beyond three years.

The model's predictions aligned closely with those of accountants, implying it considered similar financial indicators and thresholds. However, it exhibited a broader and more comprehensive view of financial health, incorporating both short-term and long-term risks. This was an area where accountants tended to be more optimistic, possibly due to the rarity of real-world bankruptcy. Another advantage of the model was its ability to process complex financial data holistically, avoiding human biases and errors in data interpretation.

While no significant difference was found between the control group and the LLM group in overall prediction performance, the LLM group showed slightly higher accuracy in edge cases, demonstrating a more critical approach to financial metrics and long-term sustainability. Additionally, accountants found the assessment process slightly easier when using the model's insights, appreciating its highlighted key financial indicators and textual summaries. However, they incorporated the model's predictions only to a limited extent, prioritizing their own judgment, consistent with existing research on AI adoption in auditing. Despite initial skepticism, all participants expressed willingness to use the model in real-world assessments, recognizing its potential as an early warning tool.

6.2 Answering the research question

As stated in Chapter 1, the following research question guided our study:

"How can LLMs be utilized and optimized effectively to enhance the going concern assessment for healthcare organizations?"

The results showed that the LLM do not currently outperform traditional auditing methods in terms of accuracy or efficiency in going concern assessments. The model did not reduce the time required for assessments and did not outperform human judgment with regards to bankruptcy prediction, leading to the rejection of Hypotheses 1 and 2. However, it still adds value by providing a critical approach to edge cases and helping to avoid human errors in data interpretation. Specifically, the model's ability to provide a more comprehensive view of financial health, by considering both short-term and long-term indicators, suggests that it can enhance accountants' judgment by offering additional structured insights that they might not typically consider. This aligns with Hypothesis 3, which was accepted, showing that accountants perceive LLMs as a useful tool to support their decision-making in going concern assessments. While the model did not significantly improve the bankruptcy prediction performance of the LLM group when compared to the control group, it did provide more comprehensive information, making the assessment process slightly easier for accountants.

The findings also highlight that LLMs are valuable in cases where long-term bankruptcy predictions are involved, with the model demonstrating an accuracy of 85% up to three years before bankruptcy. This supports its role as an early warning tool, which could assist in identifying financial distress early in the going concern assessment process.

In terms of optimization, our study emphasizes the importance of feature selection, with domain expertise proving to be a large factor to accuracy in bankruptcy prediction. The effectiveness of both prompt optimization and few-shot learning indicates that balancing these learning techniques could improve model performance further, if fine-tuned properly. However, within the scope of this research, the impact of optimization techniques remained limited, as the accuracy stayed within a range of 76% to 79%.

In conclusion, LLMs can be effectively utilized and optimized as complementary tools in the going concern assessment for healthcare organizations. While they do not replace human expertise, they can enhance decision-making by providing additional insights and early-warning capabilities. The integration of LLMs into real-world workflows could benefit accountants, especially if additional training is provided to increase confidence in their reliability. Thus, LLMs hold promise for supporting and enhancing the current process of assessing going concern of healthcare organizations.

6.3 Contributions

This research provides several key contributions to the field of auditing and bankruptcy prediction using LLMs. Firstly, it provides one of the few practical assessment of using LLMs in auditing, demonstrating their potential to improve the efficiency and accuracy of financial assessments. Secondly, it adds to the existing literature on applying LLMs specifically for bankruptcy prediction, showcasing their ability to predict business continuity risks. Moreover, this research focuses on the healthcare sector, offering valuable insights into the use of LLMs for assessing the financial stability of healthcare organizations, a context that has not been extensively explored in previous research.

Additionally, this thesis contributes to the adoption of AI within the auditing field. The findings show that LLMs can achieve accuracy comparable to that of human accountants, thus enhancing the credibility of AI in critical environments like auditing. This finding is particularly important for improving trust in AI tools among accountants who are typically skeptical about adopting AI [49].

Finally, this is one of the few studies to evaluate and compare multiple DSPy optimizers. While previous research, such as the study by Kim et al. [37], has explored this topic, most studies have focused on a single optimizer. Our comparative approach allows for a deeper understanding of the strengths and weaknesses of different optimizers, contributing to the growing literature on the DSPy framework.

6.4 Future work

Building on the findings and limitations of this research, there are several areas for future work to improve the applicability and robustness of LLM-based bankruptcy prediction models.

First of all, this thesis was conducted specifically for Dutch healthcare organizations. Given that financial structures, regulations, and auditing practices vary across industries and countries, future research should assess the model's effectiveness using more diverse organizations. Expanding the scope to different sectors and regions would provide valuable insights into the generalizability of using LLMs for going concern assessments.

Secondly, the study was conducted with a relatively small dataset and a limited number of accountants for comparison. To strengthen the reliability of the findings, future research should utilize larger datasets and a more extensive evaluation with accountants. A larger-scale study would allow for a more robust assessment of the model's accuracy and usability.

Furthermore, while DSPy was the primary framework used in this research, a comparison showed that the SAMMO framework could be a promising alternative. Further investigation into its performance in bankruptcy prediction could help determine whether it offers advantages in accuracy, efficiency, or interpretability over the DSPy framework and traditional accounting methods.

Lastly, future work can focus on more practical improvements to enhance the model's effectiveness, as outlined in the recommendations section. This includes incorporating risk assessments, optimizing prompts to reduce false negatives, integrating full auditor reports, and using multi-year financial data for better trend analysis.

Bibliography

[1] David Alaminos, Agustín Del Castillo, and Manuel Ángel Fernández. A global model for bankruptcy prediction. *PloS one*, 11(11):e0166693, 2016.

[2] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[3] Edward I Altman. Applications of distress prediction models: what have we learned after 50 years from the z-score models? *International Journal of Financial Studies*, 6(3):70, 2018.

[4] Edward I Altman and Thomas P McGough. Evaluation of a company as a going concern. *Journal of Accountancy*, 138(6):50–57, 1974.

[5] Ionuț-Florin Anica-Popa, Marinela Vrîncianu, Liana-Elena Anica-Popa, Irina-Daniela Cișmașu, and Cătălin-Georgel Tudor. Framework for integrating generative ai in developing competencies for accounting and audit professionals. *Electronics*, 13(13):2621, 2024.

[6] Kenneth Ayotte and David A Skeel Jr. Bankruptcy law as a liquidity provider. *The University of Chicago Law Review*, pages 1557–1624, 2013.

[7] BDO. Accountancy & business advice, . URL `https://www.bdo.nl/en-gb/services/accountancy-business-advice`.

[8] BDO. About - BDO, . URL `https://www.bdo.global/en-gb/about`.

[9] Rewina Bedemariam, Natalie Perez, Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Elizabeth Conjar, Ikkei Itoku, David Theil, Aman Chadha, and Naumaan Nayyar. Potential and perils of large language models as judges of unstructured textual data. *arXiv preprint arXiv:2501.08167*, 2025.

[10] Belastingdienst. Vrijstelling in de gezondheidszorg, 2023. URL `https://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/btw/tarieven_en_vrijstellingen/vrijstellingen/gezondheidszorg/vrijstelling_in_de_gezondheidszorg`.

[11] Timothy B Bell, Gary S Ribar, and Jennifer Verichio. Neural nets versus logistic regression: A comparison of each model's ability to predict commercial bank failures. 1990.

[12] Jodi L Bellovary, Don E Giacomino, and Michael D Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, pages 1–42, 2007.

[13] Big4accountingfirms. Top 10 accounting firms (2025). URL `https://big4accountingfirms.com/top-10-accounting-firms/`.

[14] International Accounting Standards Board. Ias 1: Presentation of financial statements. paragraph 26, 2021. URL `https://www.ifrs.org/standards/ias-1-presentation-of-financial-statements/`.

[15] Australian Accounting Standards Board/Auditing and Assurance Standards Board (AASB/AUASB). The impact of covid-19 on going concern and related assessments. 2020.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[17] Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, KP Subbalakshmi, and Papa Momar Ndiaye. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*, 2024.

[18] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[19] Elizabeth Carson, Neil L Fargher, Marshall A Geiger, Clive S Lennox, Kannan Raghunandan, and Marleen Willekens. Audit reporting for going-concern uncertainty: A research synthesis. *Auditing: A Journal of Practice & Theory*, 32(Supplement 1):353–384, 2013.

[20] CBS. Standard Industrial Classifications. URL `https://www.cbs.nl/en-gb/our-services/methods/classifications/activiteiten/standard-industrial-classifications`.

[21] Francesco Ciampi and Niccolò Gordini. Small enterprise default prediction modeling through artificial neural networks: an empirical analysis of i talian small enterprises. *Journal of Small Business Management*, 51(1):23–45, 2013.

[22] Welfare CIBG Ministry of Health and Sport. DigiMV. URL `https://digimv8.desan.nl/archive/search`.

[23] Pamela K Coats and L Franklin Fant. Recognizing financial distress patterns using a neural network tool. *Financial management*, pages 142–155, 1993.

[24] Denis Cormier, Michel Magnan, and Bernard Morard. The auditor's consideration of the going concern assumption: A diagnostic model. *Journal of Accounting, Auditing & Finance*, 10(2):201–222, 1995.

[25] DSPy. Examples - DSPY, . URL `https://dspy.ai/deep-dive/data-handling/examples/?h=examples`.

[26] DSPy. Optimizers, . URL `https://dspy.ai/learn/optimization/optimizers/`.

[27] DSPy. DSPy, Programming—not prompting—LMs, . URL `https://dspy.ai/`.

[28] Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A Wood. Is it all hype? chatgpt's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies*, 29(3):2318–2349, 2024.

[29] Andrea Filippo Ferraris, Davide Audrito, Luigi Di Caro, and Cristina Poncibò. The architecture of language: Understanding the mechanics behind llms. In *Cambridge Forum on AI: Law and Governance*, volume 1, page e11. Cambridge University Press, 2025.

[30] Inn Hee Gee, Peter Inho Nahm, Tieying Yu, and Albert A Cannella Jr. Not-for-profit organizations: A multi-disciplinary review and assessment from a strategic management perspective. *Journal of Management*, 49(1):237–279, 2023.

[31] Samira Ghodratnama and Mehrdad Zakershahrak. Adapting llms for efficient, personalized information retrieval: Methods and implications. In *International Conference on Service-Oriented Computing*, pages 17–26. Springer, 2023.

[32] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3, 2023.

[33] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

[34] Hussein Issa, Ting Sun, and Miklos A Vasarhelyi. Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation. *Journal of emerging technologies in accounting*, 13(2):1–20, 2016.

[35] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.

[36] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[37] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*, 2024.

[38] Hyeongjun Kim, Hoon Cho, and Doojin Ryu. Corporate default predictions using machine learning: Literature review. *Sustainability*, 12(16):6325, 2020.

[39] Tomas Krulicky and Jakub Horak. Business performance and financial health assessment through artificial intelligence. *Ekonomicko-manazerske spektrum*, 15(2):38–51, 2021.

[40] Amy Yarbrough Landry and Robert J Landry III. Factors associated with hospital bankruptcies: a political and economic framework. *Journal of Healthcare Management*, 54(4):252–271, 2009.

[41] Randall E LaSalle and Asokan Anandarajan. Auditors' views on the type of audit report issued to entities with going concern uncertainties. *Accounting Horizons*, 10(2):51, 1996.

[42] Li-lin Liu, Kathryn J Jervis, Mustafa (Mike) Z Younis, and Dana A Forgione. Hospital financial distress, recovery and closure: Managerial incentives and political costs. *Journal of Public Budgeting, Accounting & Financial Management*, 23(1):31–68, 2011.

[43] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797, 2017.

[44] Puneet Mangla. LLMOps with DSPy: Build RAG Systems Using Declarative Programming - PyImageSearch, 2024. URL `https://pyimagesearch.com/2024/09/09/llmops-with-dspy-build-rag-systems-using-declarative-programming/`.

[45] Wet marktordening gezondheidszorg, 2024. URL `https://wetten.overheid.nl/BWBR0020078/2024-01-01/#Hoofdstuk4_Paragraaf4.2_Artikel40b`.

[46] Welfare Ministry of Health and Sport. Over de jaarverantwoording, 2025. URL `https://www.jaarverantwoordingzorg.nl/over-de-jaarverantwoording`.

[47] Mahdi Moradi, Mahdi Salehi, Hadi Sadoghi Yazdi, Mohammad Ebrahim Gorgani, et al. Going concern prediction of iranian companies by using fuzzy c-means. *Open Journal of Accounting*, 1 (02):38, 2012.

[48] Mirko Moscatelli, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161:113567, 2020.

[49] Shkemb Mulliqi. Exploring the challenges and strategies of ai adoption in auditing: Insights from a big four firm. B.S. thesis, University of Twente, 2024.

[50] Nora Muñoz-Izquierdo, María Jesús Segovia-Vargas, David Pascual-Ezama, et al. Explaining the causes of business failure using audit report disclosures. *Journal of Business Research*, 98:403–414, 2019.

[51] Nora Muñoz-Izquierdo, Erkki K Laitinen, María-del-Mar Camacho-Miñano, and David Pascual-Ezama. Does audit report information improve financial distress prediction over altman's traditional z-score model? *Journal of international financial management & accounting*, 31(1):65–97, 2020.

[52] Jane F Mutchler. Auditor's perceptions of the going-concern opinion decision. *Auditing: A Journal of Practice & Theory*, 3(2), 1984.

[53] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[54] Stanford NLP. Teleprompt. URL https://github.com/stanfordnlp/dspy/tree/main/dspy/teleprompt.

[55] Stanford NLP. Modules, 2025. URL https://github.com/stanfordnlp/dspy/blob/main/docs/docs/learn/programming/modules.md.

[56] Rishi Patel. The transformation of the healthcare business through the covid-19 pandemic (2020–2021). *Journal of Risk and Financial Management*, 16(7):333, 2023.

[57] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[58] Michael E Porter. What is value in health care? *New England Journal of Medicine*, 363(26):2477–2481, 2010.

[59] Richard Riley, Bruce K Behn, and Kurt Pany. Management plans and sas no. 59 going concern resolutions. *Advances in Accounting*, 17:187–203, 2000.

[60] Muhtar Sapiri. A qualitative analysis on the role of auditors in preventing financial crises. *Golden Ratio of Auditing Research*, 4(2):89–106, 2024.

[61] Bhaskarjit Sarmah, Kriti Dutta, Anna Grigoryan, Sachin Tiwari, Stefano Pasquali, and Dhagash Mehta. A comparative study of dspy teleprompter algorithms for aligning large language models evaluation metrics to human evaluation. *arXiv preprint arXiv:2412.15298*, 2024.

[62] Tobias Schnabel and Jennifer Neville. Symbolic prompt program search: A structure-aware approach to efficient compile-time prompt optimization. *arXiv preprint arXiv:2404.02319*, 2024.

[63] Scikit-learn. Mutual info classif. URL https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html.

[64] Ravi Seethamraju and Angela Hecimovic. Adoption of artificial intelligence in auditing: An exploratory study. *Australian Journal of Management*, 48(4):780–800, 2023.

[65] Aniruddha Shrikhande. DSPy based Prompt Optimization: A Hands-On Guide, 2024. URL https://adasci.org/dspy-streamlining-llm-prompt-optimization/.

[66] Fabio Sigrist and Nicola Leuenberger. Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. *European Journal of Operational Research*, 305(3):1390–1406, 2023.

[67] Sander Steenhuis, Jeroen Struijs, Xander Koolman, Johannes Ket, and Eric Van der Hijden. Unraveling the complexity in the design and implementation of bundled payments: a scoping review of key elements from a payer's perspective. *The Milbank Quarterly*, 98(1):197–222, 2020.

[68] Koninklijke Nederlandse Beroepsorganisatie van Accountants. Accountantstitel, 2024. URL https://www.nba.nl/over-nba/lidmaatschap/accountantstitel/.

[69] Joost Wammes, Patrick Jeurissen, Gert Westert, and Marit Tanke. The dutch health care system. *International profiles of health care systems*, 137(6), 2020.

[70] J Christopher Westland and J Christopher Westland. Fundamentals of auditing financial reports. *Audit Analytics: Data Science for the Accounting Profession*, pages 1–18, 2020.

[71] Desheng Wu, Xiyuan Ma, and David L Olson. Financial distress prediction using integrated z-score and multilayer perceptron neural networks. *Decision Support Systems*, 159:113814, 2022.

[72] S Kiely Yonce and Beau Grant Barnes. Healthcare accounting research: An analysis, review, and suggestions for future work. *Journal of Governmental & Nonprofit Accounting*, 11(1):163–192, 2022.

[73] Kunpeng Yuan, Guotai Chi, Ying Zhou, and Hailei Yin. A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. *Research in International Business and Finance*, 59:101536, 2022.

Appendix A

Feature selection

Features based on expert opinion from accountants (30):
Bedrijfsresultaat, Crediteuren, Debt Service Coverage Ratio (DSCR), Eigen vermogen, Kortlopende schulden, Langlopende schulden, Liquide middelen, Liquiditeitsratio, NVTZ Omzetklasse, Overige schulden, Resultaat, Resultaatratio, Schulden aan banken, Schulden aan groepsmaatschappijen, Schulden aan kredietinstellingen, Schulden aan leveranciers en handelskredieten, Schulden aan participanten en aan maatschappijen waarin wordt deelgenomen, Schulden ter zake pensioenen, Schulden uit hoofde van financieringstekort, Schulden uit hoofde van macrobeheersinstrument, Schulden uit hoofde van subsidies, Schulden uit hoofde van te verrekenen subsidies, Solvabiliteit (debt ratio), Solvabiliteitsratio (debt ratio), Solvabiliteitsratio (weerstandsvermogen), Som der bedrijfslasten, Som der bedrijfsopbrengsten, Totaal activa, Totaal passiva, Year.

Features mutual information scores (30):
Solvabiliteitsratio (weerstandsvermogen), Solvabiliteitsratio (debt ratio), BTW-nummer, Lonen en salarissen, Debt Service Coverage Ratio (DSCR), Totaal vaste activa, Totaal activa, Rechtsvorm, STZ JN, Kostprijs van de omzet, ZKH: Psychiaters fte PUK en PAAZ (loondienst inhuur vrij beroep), Eigen vermogen, Totaal personeelsleden, Uitstroom totaal fte, waarvan zorgverleners, Totaal passiva, Opbrengsten zorgprestaties en maatschappelijke ondersteuning, Bedrijfsgebouwen en terreinen, NVTZ Omzetklasse, Aantal natuurlijke personen dat beroepsmatig zorg verleent, Totaal fte in loondienst, Resultaat na belastingen, Behoorde deze zorgaanbieder in het boekjaar tot een groep?, Uitstroom fte in loondienst, Bedrijfsresultaat, Sociale lasten, Voedingsmiddelen en hotelmatige kosten, Materiële vaste activa, Kortlopende schulden, Verantwoordingsvorm, Year.

Appendix B

Full table of model results

Table B.1 presents the complete model results. For readability, Chapter 4 displayed these results across four separate tables (Table 4.1 to 4.4). The full table is included in this appendix for completeness and to provide a comprehensive overview of all results in one place.

Table B.1: The full table of the results of the model.

Optimizer		Fold 1				Fold 2				Fold 3				Aggregated			
	Run time	Accuracy	F1	CM	FN	Accuracy	F1	CM	FN	Accuracy	F1	CM	FN	Accuracy	F1	CM	FN
Baseline (all columns)	00:23	78,57%	0,7846	[[10 4] [2 12]]	4	62,96%	0,6235	[[7 7] [3 10]]	7	77,78%	0,7778	[[10 3] [3 11]]	3	73,10%	0,7286	[[27 14] [8 33]]	14
Baseline MI-based features	00:15	78,57%	0,7846	[[10 4] [2 12]]	4	74,07%	0,7386	[[9 5] [2 11]]	5	74,07%	0,7364	[[8 5] [2 12]]	5	75,57%	0,7532	[[27 14] [6 35]]	14
Baseline expert-based features	00:13	82,14%	0,8212	[[11 3] [2 12]]	3	74,07%	0,7335	[[8 6] [1 12]]	6	77,78%	0,7778	[[10 3] [3 11]]	3	78,00%	0,7775	[[29 12] [6 35]]	12
LabeledFewShot	00:31	78,57%	0,7846	[[10 4] [2 12]]	4	74,07%	0,7335	[[8 6] [1 12]]	6	85,19%	0,8506	[[10 3] [1 13]]	3	79,28%	0,7896	[[28 13] [4 37]]	13
BootstrapFewShot	00:35	75,00%	0,7471	[[9 5] [2 12]]	5	74,07%	0,725	[[7 7] [0 13]]	7	85,19%	0,8476	[[9 4] [0 14]]	4	78,09%	0,7733	[[25 16] [2 39]]	16
BootstrapFewShot WithRandom-Search	01:26	82,14%	0,8212	[[11 3] [2 12]]	3	66,67%	0,6573	[[7 7] [2 11]]	7	81,48%	0,8148	[[11 2] [3 11]]	2	76,76%	0,7644	[[29 12] [7 34]]	12
KNNFewShot	02:14	82,14%	0,8212	[[11 3] [2 12]]	3	66,67%	0,6639	[[8 6] [3 10]]	6	77,78%	0,7759	[[9 4] [2 12]]	4	75,53%	0,7537	[[28 13] [7 34]]	13
COPRO	05:39	82,14%	0,8212	[[11 3] [2 12]]	3	74,07%	0,7386	[[9 5] [2 11]]	5	81,48%	0,8148	[[11 2] [3 11]]	2	79,23%	0,7915	[[31 10] [7 34]]	10
MIPROv2	03:25	78,57%	0,7846	[[10 4] [2 12]]	4	70,37%	0,6911	[[7 7] [1 12]]	7	85,19%	0,8519	[[11 2] [2 12]]	2	78,04%	0,7758	[[28 13] [5 36]]	13

Appendix C

Case study

C.1 Case study form

Figures C.1, C.2, and C.3 present an example of the case study, where one of the 20 assessed organizations is displayed. Both groups were asked the same question: *"Will this organization go bankrupt within two years?"* (translated from Dutch). They were also asked to explain their answer. However, they received different data to base their prediction on.

The control group received:

- One page of the auditor's report
- Financial data

The LLM group received:

- Prediction of the model
- Model-generated summary of the auditor's report page
- One page of the auditor's report
- Financial data

Pagina uit het jaarverslag

Worrell Jetten

	31-12-2022	31-12-2021
	€	€
Belastingen en premies sociale verzekeringen		
Loonheffingen	167.270	108.727
Pensioenen	44.905	26.366
	212.175	135.093
Overlopende passiva		
Vakantiegeld	19.776	10.104
Vakantiedagen	10.260	-
Accountantskosten	3.787	-
Nettoloon	15.803	9.144
Diverse overlopende passiva	150	-
	49.776	19.248

NIET IN DE BALANS OPGENOMEN ACTIVA EN VERPLICHTINGEN

Voorwaardelijke verplichtingen

Verliescompensatie

Op balansdatum is een voorwaardelijk recht op verliescompensatie niet opgenomen.

Meerjarige financiële verplichtingen

Huurverplichtingen onroerende zaken

De vennootschap is een meerjarige financiële verplichting aangegaan tot en met 31 december 2023 terzake van huur van bedrijfsruimte (€ 11.865 voor 2023 excl. servicekosten, deze zijn € 3.000 voor 2023). De huurprijs wordt jaarlijks per 1 januari op basis van de CPI-index van het CBS herzien met ingang van 1 januari 2021. De huur kan voor een aansluitende periode van 2 jaar worden voortgezet, derhalve tot en met 31 december 2025. Vervolgens kan de huurovereenkomst worden voortgezet voor aansluitende perioden van telkens 2 jaar.

Samenstellingsverklaring afgegeven d.d. 25 mei 2023 19

Financiële data

Bedrijfsresultaat: -148979.0
Eigen vermogen: -204776.0
Kortlopende schulden: 272412.0
Langlopende schulden: 93565.0
Liquide middelen: 13829.0
Liquiditeitsratio: 0.5322342628078058
NVTZ Omzetklasse: € 0 tot € 0.7 mln.
Resultaatratio: -19.954800175264623
Solvabiliteitsratio (debt ratio): 227.0314700281016
Solvabiliteitsratio (weerstandsvermogen): -33.73119282073921
Som der bedrijfslasten: 756061.0
Som der bedrijfsopbrengsten: 607082.0
Totaal activa: 161201.0
Totaal passiva: 161201.0
Year: 2022

Zal deze organisatie binnen twee jaar failliet gaan? *

- Ja
- Nee

Leg je antwoord kort uit. *

Tekst lang antwoord

Figure C.1: Example of an organization in the case study for the control group.

Model voorspelling + uitleg

Voorspelling: Ja
Uitleg: De belangrijkste financiële gegevens die wijzen op een dreigend faillissement zijn het negatieve bedrijfsresultaat van -€148.979, het negatieve eigen vermogen van -€204.776, de hoge kortlopende schulden van €272.412 tegenover de lage liquide middelen van €13.829, en de zeer lage liquiditeitsratio van 0.53. Daarnaast zijn de solvabiliteitsratio's extreem ongunstig, met een debt ratio van 227.03% en een weerstandsvermogen van -33.73%. Deze cijfers duiden op ernstige financiële problemen en een hoog risico op faillissement binnen twee jaar.

Samenvatting van pagina uit het jaarverslag

De organisatie vertoont een significante stijging in belastingen en premies sociale verzekeringen van €135.093 in 2021 naar €212.175 in 2022. Overlopende passiva nemen eveneens toe van €19.248 naar €49.776, voornamelijk door hogere vakantiegeld, vakantiedagen, en accountantskosten. De meervoudige financiële verplichtingen bestaan onder andere uit een huurverplichting van €11.865 exclusief servicekosten voor 2023. De gestegen schulden kunnen wijzen op verhoogde financiële verplichtingen die risico's vormen voor de organisatie indien deze niet zorgvuldig worden beheerd.

Pagina uit het jaarverslag

Worrell Jetten

	31-12-2022	31-12-2021
	€	€
Belastingen en premies sociale verzekeringen		
Loonheffingen	167.270	108.727
Pensioenen	44.905	26.366
	212.175	135.093
Overlopende passiva		
Vakantiegeld	19.776	10.104
Vakantiedagen	10.260	-
Accountantskosten	3.787	-
Nettoloon	15.803	9.144
Diverse overlopende passiva	150	-
	49.776	19.248

NIET IN DE BALANS OPGENOMEN ACTIVA EN VERPLICHTINGEN

Voorwaardelijke verplichtingen

Verliescompensatie

Op balansdatum is een voorwaardelijk recht op verliescompensatie niet opgenomen.

Meerjarige financiële verplichtingen

Huurverplichtingen onroerende zaken

De vennootschap is een meerjarige financiële verplichting aangegaan tot en met 31 december 2023 terzake van huur van bedrijfsruimte (€ 11.865 voor 2023 excl. servicekosten, deze zijn € 3.000 voor 2023). De huurprijs wordt jaarlijks per 1 januari op basis van de CPI-index van het CBS herzien met ingang van 1 januari 2021. De huur kan voor een aansluitende periode van 2 jaar worden voortgezet, derhalve tot en met 31 december 2025. Vervolgens kan de huurovereenkomst worden voortgezet voor aansluitende perioden van telkens 2 jaar.

Samenstellingsverklaring afgegeven d.d. 25 mei 2023 19

Figure C.2: Example of an organization in the case study for the LLM group (part 1).

Financiële data

Bedrijfsresultaat: -148979.0
Eigen vermogen: -204776.0
Kortlopende schulden: 272412.0
Langlopende schulden: 93565.0
Liquide middelen: 13829.0
Liquiditeitsratio: 0.5322342628078058
NVTZ Omzetklasse: € 0 tot € 0.7 mln.
Resultaatratio: -19.954800175264623
Solvabiliteitsratio (debt ratio): 227.0314700281016
Solvabiliteitsratio (weerstandsvermogen): -33.73119282073921
Som der bedrijfslasten: 756061.0
Som der bedrijfsopbrengsten: 607082.0
Totaal activa: 161201.0
Totaal passiva: 161201.0
Year: 2022

Zal deze organisatie binnen twee jaar failliet gaan? *

- Ja
- Nee

Leg je antwoord kort uit. *

Tekst lang antwoord

Figure C.3: Example of an organization in the case study for the LLM group (part 2).

C.2 Evaluation form

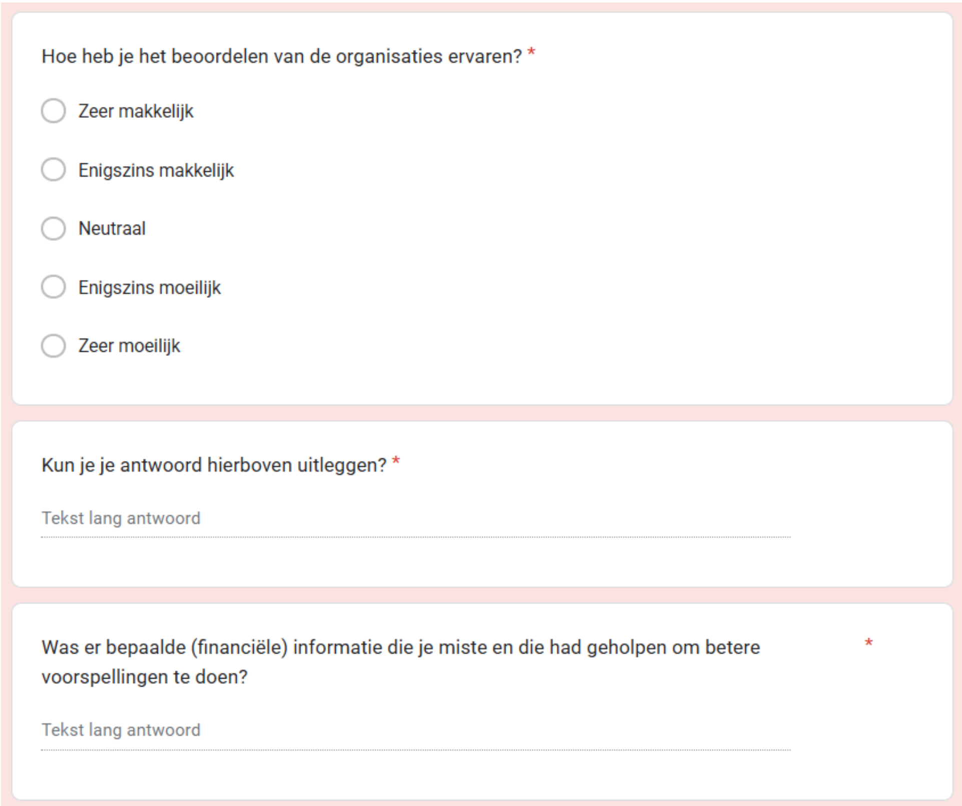
General questions for both groups, as visualized in Figure C.4:

- How did you experience assessing the organizations?
- Can you explain your answer above?
- Was there any (financial) information you missed that would have helped you make better predictions?

Model questions for the LLM group, as visualized in Figure C.5:

- To what extent did you take the model's predictions into account in your own prediction?
- Can you explain your answer above?
- Which aspects of the model did you find most valuable or insightful for your own prediction, and why? (None is also an answer.)
- Would you use the model's predictions in real-world assessments?
- Do you have any other feedback on the model?

Tables C.1 to C.4 present the answers to these questions from all participants.


Hoe heb je het beoordelen van de organisaties ervaren? *

- Zeer makkelijk
- Enigszins makkelijk
- Neutraal
- Enigszins moeilijk
- Zeer moeilijk

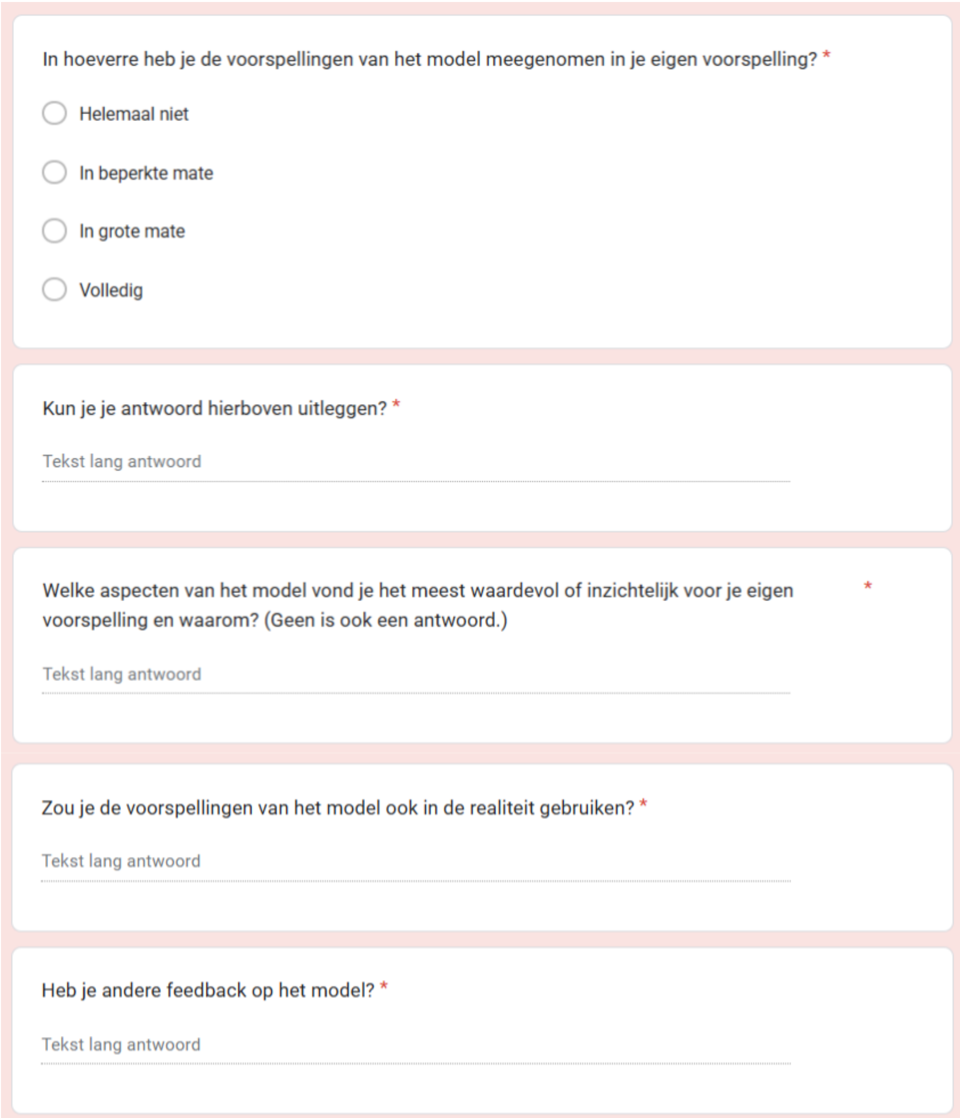
Kun je je antwoord hierboven uitleggen? *

Tekst lang antwoord

Was er bepaalde (financiële) informatie die je miste en die had geholpen om betere voorspellingen te doen? *

Tekst lang antwoord

Figure C.4: Evaluation form with general questions for both groups.



In hoeverre heb je de voorspellingen van het model meegenomen in je eigen voorspelling? *

- Helemaal niet
- In beperkte mate
- In grote mate
- Volledig

Kun je je antwoord hierboven uitleggen? *

Tekst lang antwoord

Welke aspecten van het model vond je het meest waardevol of inzichtelijk voor je eigen voorspelling en waarom? (Geen is ook een antwoord.) *

Tekst lang antwoord

Zou je de voorspellingen van het model ook in de realiteit gebruiken? *

Tekst lang antwoord

Heb je andere feedback op het model? *

Tekst lang antwoord

Figure C.5: Evaluation form with model questions for the LLM group.

Table C.1: Ease of evaluating for control group.

Score	**Key Theme**	**Explanation**
Very Easy	—	—
Somewhat Easy	Data reliability	Based on the ratios and given information, it was easy to determine the current financial performance. Unfortunately, there were many errors in the data, which made me doubt the reliability of the information. Evaluating an organization's performance purely based on numbers is only one step; however, the story behind the organization (reasons for existence, added value, shareholders, etc.) is wrongly left out of the evaluation.
Somewhat Easy	Financial Ratios	Mainly focused on financial ratios. This was easy to follow.
Neutral	Ratio explanation and readability	Sometimes it was unclear how certain ratios were explained, making it a guess whether something was interpreted positively or negatively. And having dots between large numbers is always helpful so you don't have to think about how big the number is.
Somewhat Difficult	Limited data	Financial data was mostly available for only one year, making it difficult to assess whether there were incidental setbacks or not. Sometimes this could be inferred from the given equity.
Somewhat Difficult	Missing key information	A lot of necessary information was missing, which is crucial for better assessing whether an organization might go bankrupt, such as: the presence or absence of bank covenants and any breaches, liquidity forecasts, management's financial expectations, future plans, etc.
Very Difficult	—	—

Table C.2: Ease of evaluating for LLM group.

Score	**Key Theme**	**Explanation**
Very Easy	—	—
Somewhat Easy	Financial data clarity	Clear financial data. The text from the model is sometimes harder to assess for its value. Not everything included is always relevant.
Somewhat Easy	Financial data sufficiency	I still had the feeling that I wanted to see more of the full financial statements, even though the available information was basically sufficient.
Somewhat Easy	Key figures and no complex aspects	Key financial figures were displayed with a summary. Cases were generally quite clear. Complex aspects (liquidity budgets, bank covenants, letter of support from the parent company) were left out of consideration.
Neutral	Lack of overall view	Sometimes, a bit of an overall view of the balance sheet and P/L statement is missing. However, there was a clear distinction between high and low risk due to negative operating results and negative equity.
Neutral	Limited data for multiple time periods	Information was often limited to two years, meaning that historical data and future predictions were missing. In some cases, that information can be decisive.
Somewhat Difficult	—	—
Very Difficult	—	—

Table C.3: Information that accountants felt was missing.

Group	Key Themes	Responses from accountants (translated from Dutch)
Control group	Explanation of ratio calculations	Sometimes, an explanation of how a ratio was calculated would have been helpful so that you don't have to think about it within the 45 minutes available. Ratios can be calculated in many different ways, and I occasionally had to spend time figuring this out, which took up more time than I had to complete everything in 45 minutes.
	Additional explanatory notes	The attached explanations.
	Comparative figures from previous years	Comparative figures from previous years.
	Continuity section from the annual report	The continuity section in the financial statements is always included, so I would add it. This provides insight into how management views continuity (and I think it also clarifies the points I mentioned in the previous question).
	Revenue	The revenue.
LLM group	More historical data for trend analysis	With multiple years or future figures, the model could potentially become more accurate.
	Operational cash flow and incidental financial events	Operational cash flow for the year and any information from the income statement that indicates incidental income and expenses.
	More structured response	A slightly more structured response.
	Management insights and strategic plans	Yes, forecasts and information from the management of the healthcare institution, such as possibilities for obtaining financing (and the supporting rationale), bank covenants, letters of support, agreements with insurers/clients, reorganizations, etc. Right now, the model mainly looks at information solely from the financial statements.
	Overall financial position	Balance sheet and income statement provide a more complete picture and indicate possible shifts between items.

Table C.4: Key aspects of the model part of the evaluation form (LLM group).

Question	**Key Themes**	**Responses from accountants (translated from Dutch)**
To what extent did you take the model's predictions into account in your own prediction? Explain your answer.	Initial independent judgment before checking model output	To a limited extent - Mostly formed my own judgment first and then looked at the prediction. Additionally, the outcome based on the ratios was often clear.
	Model as supporting tool	To a limited extent - I did make my own assessment after reading the model's output. The model provided a lot of guidance, which made me rely on it more as I progressed through the test. However, I deviated from the model's outcomes twice.
	As first impression and then deeper analysis	To a limited extent - It does provide an initial trend and insight, but ultimately, you still look deeper into the financial data than the model does.
	Attempt to minimize reliance on model	To a limited extent - Tried to leave it somewhat aside.
	Strong reliance on model insights	To a large extent - The model's prediction mainly provided a helpful summary in which ratios were also calculated. These were useful and often confirmed the overall picture.
Which aspects of the model did you find most valuable or insightful for your own prediction, and why? (None is also an answer.)	Liquidity assessment and broader financial insights	The liquidity ratio, as it indicates whether the company can meet its short-term payment obligations. Additionally, it was useful that the model also considered non-balance sheet commitments (such as lease obligations) and guarantees, as well as trends in the P&L, which are not immediately apparent from the ratios and balance sheet.
	Key financial indicators	Equity / result / liquidity.
	Text-based summary and numerical insights	A summary of the considerations in text form and a listing of relevant amounts/percentages.
	Calculated financial ratios	The calculated ratios.
	Highlighting weak financial ratios	The highlights of the weak financial ratios.
Would you use the model's predictions in real-world assessments?	Yes, if integrated into workflow with minimal effort	If it requires little to no effort to use, then yes. It would work well combined with a specific dashboard.
	Yes, useful tool	Yes, useful.
	Yes, for early warning and initial assessment	Yes, it provides a good early warning, and I would like to use it as the first evaluation in the file.
	Yes, for summarizing financial data	Yes, mainly for a good summary of the figures.
	Yes, as a supplement to own analysis	Yes, as a supplement to my own analysis.
Do you have any other feedback on the model?	Remove unnecessary information	Perhaps remove the imported page and focus only on balance sheet and P&L data. The NIBOVs are too inconsistently filled in.
	Irrelevant risk classifications	It keeps selecting the risk related to macro control instruments, even though that is not necessarily relevant.
	Structured risk assessment with clear categorization	A few bullet points where we first identify a risk as low / medium / high, followed by further details.
	No further feedback	No, I mainly miss the considerations regarding additional information from the institution, as mentioned above. No