CRC Credit Research Centre

CRC Working Papers

Unleashing the power of text for credit default prediction: Comparing human-generated and AI-generated texts

By Zongxiao Wu, Yizhe Dong, Yaoyiran Lib, Baofeng Shic

March 2024

Keywords: Bank lending, Large language model, Credit risk, Text mining, Deep learning, ChatGPT

THE UNIVERSITY of EDINBURGH

Contact
Credit Research Centre
University of Edinburgh Business School
29 Buccleuch Place
Edinburgh, EH8 9JS
credit.research.centre@ed.ac.uk

Unleashing the power of text for credit default prediction: Comparing human-generated and AI-generated texts

Zongxiao Wu[a], Yizhe Dong[a *], Yaoyiran Li[b], Baofeng Shi[c]

[a] Business School, University of Edinburgh, Edinburgh, EH8 9JS, United Kingdom

[b] Language Technology Lab, University of Cambridge, Cambridge, CB3 9DA, United Kingdom

[c] College of Economics and Management, Northwest A&F University, Yangling, 712100, China

Abstract

Despite the increasing use of generative large language models (LLMs) across various domains, their potential applications in the financial sector remain largely unexplored. This study pioneers the exploration of integrating ChatGPT, a prominent LLM, into lending decision-making with a focus on leveraging text data for credit default prediction. We first utilise ChatGPT to analyse and interpret textual loan assessments originally generated by loan officers, which enables it to produce human-like analyses. Comparing these artificial intelligence (AI)-generated texts with the original human-generated texts, we observe a number of noticeable distinctions between the two, particularly regarding text length, semantic similarity, and linguistic representations. By employing deep learning techniques, we find that, along with conventional structured data, the inclusion of unstructured text data, particularly ChatGPT-generated texts, can significantly improve credit default predictions. We further demonstrate how the contents of both human-generated and ChatGPT-generated loan assessments contribute to the models' prediction and find that the effect of essential words on the prediction results is highly context-dependent. Finally, we find that ChatGPT's analysis of the borrower's delinquency is the most influential factor for predicting eventual default. Our work provides some of the first insights into the potential applications of generative AI for financial and business services.

Keywords: Bank lending; Large language model; Credit risk; Text mining; Deep learning; ChatGPT

* Corresponding Author.
Email addresses:
Zongxiao.Wu@ed.ac.uk, Yizhe.Dong@ed.ac.uk, yl711@cam.ac.uk, shibaofeng@nwsuaf.edu.cn

"*Although humans are still better than GPT at a lot of things, there are many jobs where these capabilities are not used much. […] and soon these data sets will also be used to train the AIs that will empower people to do this work more efficiently*".

-- *The Age of AI has Begun* by Bill Gates[1]

1. Introduction

Since its launch in November 2022, OpenAI's ChatGPT, powered by a large language model (LLM), now boasts over 150 million users. This success has triggered a global wave of innovation in the field of artificial intelligence (AI). The unprecedented advancements in generative AI are reshaping industries and revolutionising the way businesses operate and information acquired in the digital era. This technology can generate novel content, including text, images, audio, codes, simulations, and synthetic data, by analysing patterns and structures within existing data. We have seen an explosion in generative AI applications for software, marketing, gaming, e-commerce, healthcare, and education companies. However, its potential within the financial sector remains largely untapped. Financial institutions and investors still lack sufficient knowledge and understanding to make informed decisions regarding the incorporation of generative AI into investment and lending strategies. Therefore, this paper aims to bridge these practical challenges by studying the adoption of generative AI tools and natural language processing (NLP) techniques in the lending decision-making process and thus to show the importance of leveraging alternative data to supplement traditional credit data.

Over the last decade, machine learning (ML) has played an instrumental role in propelling the lending industry forward, enabling notable advancements such as improved default prediction, better underwriting, and greater financial inclusion. Many studies have adopted ML techniques such as neural networks (Akkoç, 2012), decision tree (DT) (Xia et al., 2017), random forests (RF) (Jiang et al., 2019), XGBoost (XGB) (Gunnarsson et al., 2021) and support vector machine (SVM) (Maldonado et al., 2017) for default prediction and have shown that they are superior to traditional statistical methods. In the era of big data, several recent studies have sought to leverage text data, a major source of alternative data, for credit default prediction (e.g., Jiang et al., 2018; Stevenson et al., 2021; Kriebel & Stitz, 2022; Liu et al., 2023). They utilise NLP tools such as word embedding models, convolutional neural networks (CNNs) and transformer models to extract insights from text data and demonstrate that integrating this unstructured data into the modelling process can significantly improve predictive performance. However, most of these studies have centred around peer-to-peer lending. There is limited evidence pertaining to traditional bank loans and exploiting the recent advances from LLMs in financial and business contexts.

[1] *The Age of AI has begun* by Bill Gates. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

To the best of our knowledge, this paper is the first to assess the capabilities of ChatGPT in analysing and distilling information from loan-related text within the context of credit default prediction. Specifically, we endeavour to examine the following research questions:

RQ1. What commonalities and differences can be discerned between the texts generated by ChatGPT and those written by human beings?

RQ2. To what extent can textual loan assessments be leveraged to predict borrowers' defaults? If applicable, which type of text exhibits higher effectiveness, human-generated or AI-generated?

RQ3. Which particular signal words and phrases within these two types of texts strongly contribute to predictive capacity?

RQ4. Do all sections of AI-generated text carry equal potential for predicting defaults? If not, which section yields the most predictive capacity?

Our empirical analyses are based on a unique loan dataset containing 2,460 micro and small business (mSE) loans which contain both structured and text data. We first utilise ChatGPT to analyse the original human-written loan assessments and then derive AI-refined analytical assessments. Through comparisons between these two types of texts, we find that human-generated and ChatGPT-generated texts are quite different in terms of text length, semantic similarity, and linguistic representations. Then, based on state-of-the-art NLP models, we develop text-enhanced credit scoring models which leverage human-generated and ChatGPT-generated texts, respectively. While both texts benefit credit scoring models (compared to models using only structured data), ChatGPT-generated texts can lead to more accurate prediction results. Furthermore, we apply the local interpretable model-agnostic explanations (LIME) approach to explore which words/phrases lead to an improved result and find that the effect of essential words on prediction results depends on the context in which the word occurs. Moreover, our extensive analysis indicates that ChatGPT-generated texts are notably more effective in improving default predictions, possibly due to ChatGPT's robust capability to identify the borrower's delinquency factors by assessing the information contained in the original loan assessments. Finally, our further analysis compares the efficacy of several alternative generative AI models with ChatGPT: we find that OpenAI's generative pre-training transformer (GPT)-based LLMs generally outperform other open-source LLMs in loan default prediction.

Our study makes several contributions to the field. First, our work benefits the fast-growing research on generative AI applications in finance and business alike. While previous works have demonstrated that ChatGPT can facilitate stock return prediction (Lopez-Lira & Tang, 2023; Kim et al., 2023), our paper is the first to examine the potential of generative AI to enhance lending decision-

making and credit risk mitigation. Our results suggest that AI-generated texts add considerable value to loan default prediction as they can amplify useful information in loan officers' textual assessments. Our paper also compares the efficacy of multiple generative AI models, which provides actionable insights for institutions and organisations that intend to embrace generative AI tools to improve their financial decision-making. Second, our study complements and contributes to the existing literature on the value of alternative data, especially AI-generated text data, for predicting important outcomes in financial markets, including voice (Yang et al., 2023), visual content (Bazley et al., 2021), Google search data (Da et al., 2015), and satellite data (Mukherjee et al., 2021). Our research demonstrates that including such texts can significantly improve loan default prediction for mSEs whose structured data (i.e., financial information) are often unavailable to lenders. Third, our research is the first to comprehensively investigate the linguistic differences between loan assessment texts generated by humans and LLMs. Our results indicate that these two types of texts are significantly different concerning text length, semantic similarity, and linguistic representations.

The remainder of the paper is organised as follows. In Section 2, we discuss prior literature on NLP-based default prediction and recent LLM applications in finance and business sectors. Section 3 provides a detailed description of our dataset and the ChatGPT prompt proposed for generating new content. Section 4 outlines the modelling process and methods employed in this study. In Section 5, we investigate the linguistic differences between human-generated and ChatGPT-generated texts, present our prediction results and analyse how text influences the model performance (respectively corresponding to our four research questions). We also provide further analysis of the default prediction experiments using texts generated by alternative generative AI tools. Finally, we summarise our work's key findings and implications in Section 6.

2. Literature review

Here, we organise the prior literature into two main categories and point out the research gaps that motivate our work. Section 2.1 briefly summarises the use of NLP techniques in credit default prediction; Section 2.2 reviews recent studies exploring LLMs' potential in finance and business sectors.

2.1 NLP-based default prediction

Credit scoring refers to an automatic credit assessment procedure that helps lenders identify borrowers who will fail to fulfil their financial obligations within a group of loan applicants (Li et al., 2017). Given the surge in the consumer lending market, novel credit scoring models continue to attract considerable interest in academia and industry (Song et al., 2023). Although traditional credit scoring models primarily relied on structured data, strong evidence shows that integrating text data can

significantly enhance prediction results by leveraging modern NLP techniques (Wang et al., 2020; Kriebel & Stitz, 2022).

Early credit scoring studies explored the value of text by leveraging traditional text mining approaches. For example, Herzenstein et al. (2011) highlighted the significant influence of identity claims such as 'trustworthy', 'successful', and 'moral' extracted from loan narratives on lending decisions. Iyer et al. (2016) derived several characteristics from borrowers' loan descriptions, including average word length and percentage of misspelt words, among others, to predict loan default. Although these approaches can to some extent exploit the relationship between texts and human cognitive processes, their application requires expert knowledge and significant manual efforts (Kriebel & Stitz, 2022). As digital technologies flourished, scholars started to employ statistical NLP models to extract textual features. Pioneering these approaches, Jiang et al. (2018) used latent Dirichlet allocation (LDA), a widely used topic model, to derive features from borrowers' written descriptions. Fitzpatrick & Mues (2021) applied a biterm topic model to generate text representations for loan descriptions, thereby addressing the sparsity problem of short texts. Other statistical methods, including latent semantic analysis and term frequency – inverse document frequency (TF-IDF), have also been widely used in credit scoring studies (Netzer et al., 2019; Xia et al., 2020). However, while these statistical techniques are easy to scale, they may fail to comprehend complex linguistic phenomena and suffer from the curse of dimensionality.

Given the breakthrough in deep learning on various fields, scholars have increasingly explored the potential of neural NLP models for extracting textual information. Earlier work adopted static word representations such as global vectors for word representation (GloVe) and fastText to transform words into real-valued vector representations which can be used for downstream tasks with subsequent classifiers. In the context of default prediction, Mai et al. (2019) pioneered the use of CNNs, while Matin et al. (2019) simultaneously leveraged recurrent neural networks to predict corporate bankruptcies, gleaning insights from audit reports and management statements; Wang et al. (2020) and Kriebel & Stitz (2022), however, adopted more traditional classifiers such as logistic regression (LR) and RF to predict borrower's default by leveraging GloVe embeddings. More recently, pretrained language models have revolutionised the field of NLP, offering notable performance enhancements: neural networks are initially pre-trained on massive amounts of human language data and then adapted to various downstream tasks by task-specific fine-tuning (a way of transfer learning). Transformer models, such as bidirectional encoder representations from transformers (BERT) and the robustly optimised BERT pre-training approach (RoBERTa), are now state-of-the-art across many benchmark NLP tasks. While these

models are well-known in NLP and AI communities, it was not until 2021 that they were introduced for credit default prediction. Stevenson et al. (2021) shed the first light on utilising BERT to derive text representations from loan officers' assessments for bankruptcy prediction. Subsequently, Fitzpatrick & Mues (2021) applied BERT, embeddings from language models (ELMo) and the universal sentence encoder (USE) to capture credit-relevant information contained in borrowers' loan descriptions; with BERT and RoBERTa, Kriebel & Stitz (2022) showed that even succinct text descriptions of borrowers can considerably enhance default prediction.

2.2 LLMs' applications in finance

The emergence of LLMs can be traced back to the development of transformer architecture that exploited the attention mechanism to model long-range dependencies in a sequence (Vaswani et al., 2017; Susarla et al., 2023). Transformer-based LLMs with huge number of parameters (e.g., billions) are trained on massive text data from diverse sources, and they can be adapted to a broad spectrum of downstream tasks. Evidence suggests that LLMs achieve impressive results over a range of NLP tasks, including logical reasoning, question answering, and code generation (Wu et al., 2023). In November 2022, the release of OpenAI's ChatGPT catapulted LLMs to the centre of the NLP community and attracted a remarkable amount of discussion worldwide. Building upon existing GPT infrastructure, ChatGPT fuses unsupervised learning and reinforcement learning derived from human feedback, generating responses to inquiries that closely mimic human discourse. Given the unprecedented capabilities of LLMs such as ChatGPT, some studies have started to explore their potential usage in various domains, but there are few relevant studies in the financial field (Lopez-Lira & Tang, 2023).

Most of the existing literature provides guidelines and potential application scenarios for the use of LLMs within the financial domain, taking ChatGPT as a core example. Concerning deployment of LLMs on concrete tasks, Wenzlaff & Spaeth (2022) first explored how ChatGPT defines concepts such as crowdfunding, alternative finance, and community finance, and how these concepts correspond to human answers in academic scholarship. Dowling & Lucey (2023) suggested that ChatGPT could generate plausible-seeming financial research studies, and the results could be significantly improved by integrating private data and researcher expertise. Cao & Zhai (2023) presented several examples on how to guide ChatGPT to generate outcomes for financial and accounting research questions, including sentiment analysis, ESG (environmental, social, and governance) analysis, corporate culture analysis, and Federal Reserve opinion analysis. Alshurafat (2023) reviewed the potential use of ChatGPT in financial and accounting applications and also discussed the ethical considerations regarding its use.

Other literature has delved into using LLMs in financial market applications leveraging quantitative

methodologies, especially in equity investment. For example, Lopez-Lira & Tang (2023) examined the capability of various LLMs for predicting stock returns via sentiment analysis of news headlines. Their findings suggest that ChatGPT outperforms traditional sentiment analysis models, and integrating such advanced LLMs into investment decision-making could lead to more accurate prediction results and enhance the performance of trading strategies. Yue et al. (2023) applied an explainable AI model to ChatGPT in order to interpret the stock performance generated by nonlinear machine learning models. In another study, Ko & Lee (2023) demonstrated the efficacy of ChatGPT on portfolio management, particularly in asset allocation and diversification. Their analysis shows that the portfolios selected by ChatGPT are notably superior, in terms of a diversity index, to randomly selected assets. Additionally, Kim et al. (2023) used ChatGPT to distil textual information disclosed in companies' annual reports and conference calls and found that the ChatGPT generated content is more effective at explaining stock market reactions.

In summary, our review of existing studies reveals the importance and effectiveness of using NLP techniques to extract text features for default prediction, and there is indeed growing interest in applying LLMs to address financial problems. However, to the best of our knowledge, no existing research has yet examined how to incorporate LLMs into making informed decisions that will supplement lending strategies and enhance credit risk mitigation. In this paper, we strive to bridge these research gaps, which we believe could provide insights for broader financial and business applications.

3. Data description and collection

The mSEs loan dataset used in this study was provided by a Chinese bank operating on a national basis. After removing records with apparent errors, the dataset consists of 2,460 loans with durations ranging from 1 to 8 months, originating between December 2008 and July 2009. All the loans are closed – 2,400 loans have been fully paid off at maturity or early, and 60 borrowers defaulted.[2] Default is defined as 90 days past payment due. Each borrower has 19 nominal and 12 numeric attributes, one text extracted from the loan assessments generated by a loan officer and a final class label (i.e., defaulter or non-defaulter). Besides the original text, we also derived a new text from the ChatGPT-generated analytical assessments.

3.1 Structured features

The original dataset comprises 31 standard credit features, including information on borrowers'

[2] We note that the dataset is characterised by a severely imbalanced ratio of 1:40 (defaulter: non-defaulter). However, this was quite common among Chinese lenders during this time and remains so today. For example, according to the China Banking and Insurance Regulatory Commission, the average non-performing loan ratio for the Chinese banking sector was 1.58% in 2009.

demographics, business details and loan particulars. In cases where continuous features are missing, they are filled with the means of features, while missing values for categorical features are replaced with new categories. Appendix Table A.1 presents the definitions and summary statistics of all standard variables in our dataset.

To ensure consistency and comparability of input features across models, we apply the weight of evidence (WoE) method on the features beforehand (Jiang et al., 2019). The WoE value in each category/bin for a given feature is calculated as the logarithm of the proportion of non-defaulters to the proportion of defaulters in that particular category/bin. A large negative value corresponds to a higher default risk and vice versa (Yap et al., 2011). We also employ information value (IV) and variance inflation factor (VIF) measures to remove redundancy, reduce the multicollinearity of input features and select an appropriate subset of features for inclusion in our credit scoring models. We calculate the IV as the weighted sum of the WoE value for each category/bin of the feature, which serves to represent the feature's predictive power. In this study, we retain those features with an IV greater than 0.01 and less than 0.50 (Anderson, 2007). Meanwhile, we use the VIF to detect multicollinearity issues among the features and only retain features with a VIF less than or equal to 10. Through these preprocessing steps, we identify 18 structured variables for our subsequent modelling.

3.2 Human-generated text

In many cases, mSEs lack a credit history and struggle to provide necessary and reliable financial information. To obtain an all-encompassing view of an mSE's operation, loan officers from the lender usually conduct site visits. These visits serve to verify the accuracy of the information provided in the loan applications and are a way to gather first-hand information about the borrower's business, day-to-day operations, and willingness to repay, among others. Subsequently, based on the information gathered, the loan officers prepare loan assessment reports, which assist the lenders in assessing the potential credit risk of the borrowers and in making informed decisions on loan approvals. Within our dataset, the loan assessment text generally evaluates the borrower from three essential perspectives: an overall impression of the borrower, the borrower's repayment intentions, and the borrower's assets and liabilities. To ensure that the text is clear for our analysis, we remove redundant white spaces in the text and add a full stop at the end of each complete sentence, steps which can enhance the text's readability. Below, we present one typical extract from the defaulters' textual loan assessments, initially in Chinese, which has been translated into English. This extract has been fully anonymised, with all quantitative values masked. For comparison, a typical extract from non-defaulters' textual loan assessments is available in Appendix B.

Human-generated text for a defaulter: *"The borrower has good peer relationships, actively*

cooperates with the credit officer's investigation, and provides valid documentation. The borrower has marketing work experience and a certain understanding of the industry's business environment. The borrower knows the consequences of loan default, and no obvious factors affect his willingness to repay. However, through an individual credit check, the credit officer found that the borrower had a history of several late credit card payments caused by his being busy with work and not making timely payments, thus reducing his credit rating to A. The borrower's cash and bank balance were approximately ¥ ???, and the borrower's current receivables were approximately ¥ ???, which is consistent with the borrower's statements. The inventory in the borrower's shop is valued at the purchase price of approximately ¥ ???."

As illustrated in Figure 1, the text lengths for both the Goods (non-defaulters) group and Bads (defaulters) group seem to roughly follow symmetrical Gaussian distributions. However, there are notable differences between the two groups. The Bads group exhibits a lower mean word count of 164 and a larger standard deviation (SD) of 93. In contrast, the Goods group displays a higher mean word count of 210 and a smaller dispersion (SD 75). These findings suggest that loan officers tend to generate more extended and detailed texts for loans deemed safe, while providing more concise assessments for risky loans. One possible explanation for this pattern is that non-defaulters tend to provide more positive attributes, while both non-defaulters and defaulters downplay or conceal potential risk factors. When borrowers emphasise their positive attributes, this may lead loan officers to focus more on the strengths of the loan application, resulting in more extended assessments. Conversely, when borrowers provide little information, loan officers may face difficulties uncovering hidden risk factors, which can contribute to more concise loan assessments. By recognising the tendencies in the text length and levels of detail in loan assessments, we can gain insights into the assessment practices of loan officers and obtain potential implications for predicting default risk.

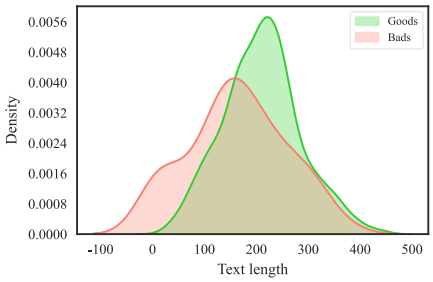


Figure 1 Text length of human-generated texts

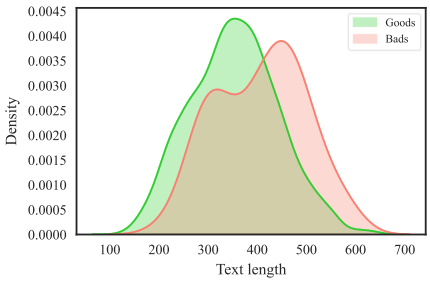


Figure 2 Text length of ChatGPT-generated texts

3.3 ChatGPT-generated text

In this study, we employ ChatGPT[3] to analyse the human-produced loan assessments and generate

[3] The version of ChatGPT used in this paper is version 2.13, based on the standard ChatGPT model, as updated on 13th February 2023.

AI-refined analytical assessments. To enhance ChatGPT's understanding of the task and instruct it to generate the desired responses, we adopt a well-designed prompt template which is set to be clear and specific, containing sufficient background. Our prompt template consists of two main components: (1) the initial loan assessment about the borrower provided by the loan officer, (2) an outline of the task that we want ChatGPT to perform. Specifically, we instruct ChatGPT to identify factors that support a borrower's likelihood of repayment and factors that may lead to the borrower's default. These are crucial factors in the lender's decision-making process. Note that we do not impose any constraints on the ChatGPT's output length. Since ChatGPT was only available through its website in February 2023, we obtain the responses by entering the prompts into the input box. The chat history can influence the answers produced by ChatGPT to a large extent, so we refreshed the thread for each question. Below, we provide the detailed prompt template employed in this study for each original human-generated loan assessment:

Prompt template: *"Hi ChatGPT, there is a bank loan borrower whose details are below." + [Insert the original texts] + "The bank plans to lend to this borrower. Based on the above information, please carefully summarise and analyse the factors that support the borrower's ability to repay the loan on time and the factors that could lead to the borrower's default. The expected answer template consists of two parts: 1. Factors supporting the borrower's repayment: [Insert answer here]; 2. Factors that could potentially lead to the borrower's default: [Insert answer here]."*

We feed the templated text to ChatGPT and retain ChatGPT's analysis of the factors that favour the borrower's repayment and those that have the potential to lead to the borrower's default. The following is the output of ChatGPT corresponding to the previous example.

ChatGPT-generated text for a defaulter: *"1. Factors supporting the borrower's repayment: * The borrower has good peer relationships, actively cooperates with the credit officer's investigation, and provides valid documents, which indicates that the borrower has a good cooperative attitude and integrity. This is conducive to establishing a cooperative relationship with the bank. * The borrower has a stable social status and some resources, which may have a positive influence on the borrower's future repayment. * The borrower attaches importance to the risk of default and has no obvious factors affecting his willingness to repay, which indicates that the borrower has the willingness and ability to repay, and this may help the bank to conduct risk assessment and control. 2. Factors that could lead to the borrower's default: * The borrower's personal credit check shows that there were several overdue payments on credit cards, and this leads to a downgraded credit rating of A. Although the borrower indicates that he is usually busy with work and has not made timely payments, this may have a negative impact on the borrower's credit risk assessment. Therefore, the bank needs to carefully consider the setting of the loan amount and interest rate, etc. * The borrower has a large amount of receivables, which may affect the borrower's future cash flow and hence his ability to repay. Therefore, banks need to conduct a detailed review of the borrower's receivables to fully examine his true business operation*

and consider the impact of this factor in risk control."

Figure 2 shows the text length distributions for the Goods and Bads groups' ChatGPT-generated texts. The text lengths of the Goods group appear to follow a Gaussian distribution, with a mean value of 354 (SD 89), while those of the Bads group exhibit a bimodal distribution, with a mean value of 406 (SD 90). These results contrast with our earlier findings obtained from human assessments, suggesting that ChatGPT tends to produce more detailed and extended assessments for borrowers who are considered riskier. Moreover, upon examining the contents of human-generated and ChatGPT-generated texts, we observe that ChatGPT not only provides a good summary of the original information but also offers additional extended analyses and implications regarding the creditworthiness of borrowers. The results unveil ChatGPT's impressive capability to summarise information while simultaneously enriching content.

4. Experimental design

4.1 Model specifications

The main objective of our study is to compare the predictive capabilities of human-generated and AI-generated texts by leveraging NLP approaches for credit risk analysis. Specifically, we deploy four representative NLP techniques, namely LDA, fastText, Ada-002, and BERT, to derive numerical representations from texts that can be utilised for subsequent credit scoring models. To facilitate comparisons among techniques, we use the multi-layer perceptron (MLP) neural network as the benchmark classifier to predict default (Stevenson et al., 2021). Our prediction models can be built based on three types of features: structured (comprising only structured features), text (comprising only text features) and combined (a combination of both structured and text features). Thus, we end up with 17 different models, as presented in Table 1.

Table 1 Input features for various default prediction models

	MLP			
Structured	7 continuous 11 categorial			
	LDA+MLP	fastText+MLP	Ada-002+MLP	BERT+MLP
Text-only (Human)	30 topics	300-dim vector	1536-dim vector	768-dim vector
Combined (Human)	7 continuous 11 categorical 30 topics	7 continuous 11 categorical 300-dim vector	7 continuous 11 categorical 1536-dim vector	7 continuous 11 categorical 768-dim vector
Text-only (ChatGPT)	30 topics	300-dim vector	1536-dim vector	768-dim vector
Combined (ChatGPT)	7 continuous 11 categorical 30 topics	7 continuous 11 categorical 300-dim vector	7 continuous 11 categorical 1536-dim vector	7 continuous 11 categorical 768-dim vector

LDA is an unsupervised theme model used to find natural groups of topics, which models each

document by identifying the underlying themes, represented as a probability distribution over a certain set of words (Blei et al., 2003). LDA has a three-layer Bayesian structure consisting of documents, topics and words, and its posterior distribution depends on the parameters in the previous layer. The most crucial tuning parameter for LDA that needs to be predefined is the number of topics, denoted by n. In this study, we train the LDA model on the training set and perform hyperparameter searches within the range of [5, 10, 15, 20, 25, 30] to determine the best n for the LDA model based on MLP's prediction results on validation samples. These searches imply that LDA models with 30 topics should be utilised in our subsequent modelling.

Besides, we also derive pre-trained fastText word embeddings from two types of texts. fastText, developed by Meta AI, is trained with the efficient hierarchical Softmax and skip-gram, which primarily focuses on achieving scalable solutions for text classification and representation learning while processing large datasets quickly and accurately. Compared to other earlier works such as word2vec, fastText features a deeper understanding of the internal structure of words by modelling each word as a bag of character n-grams. Due to its ability to capture a word's morphology and sub-word information, fastText can also provide vector representations for out-of-vocabulary words. Following Bojanowski et al. (2017), we map text inputs to 300-dimensional vectors using pre-trained fastText embeddings, which are then utilised in our modelling process.

Moreover, we also include OpenAI's latest embedding model, Ada-002, to convert our texts into numerical representations. Ada-002, known as text-embedding-ada-002, is a transformer-based text representation model proposed by OpenAI in December 2022 that can produce sequence embeddings for text inputs. It outperforms all of OpenAI's previous embedding models on text search, code search, and sentence similarity tasks and achieves a comparable performance on text classification. As in the use of fastText, we use pre-trained Ada-002 embeddings to generate a 1536-dimensional vector for each text input that can be utilised in our subsequent credit scoring models.

Finally, following Stevenson et al. (2021), Fitzpatrick & Mues (2021) and Kriebel & Stitz (2022), we also deploy BERT to extract contextualised representations for our texts. BERT is a transformer model proposed by Google AI (Devlin et al., 2018) which is trained with masked language modelling and next sentence prediction objectives. BERT contains a series of so-called transformer blocks, each consisting of a bidirectional self-attention mechanism followed by feed-forward layers. Instead of reading the text input sequentially, the bidirectional paradigm allows an entire sequence of words to 'attend' to each other, thus enabling BERT to learn the word meanings in context. BERT exhibits remarkable adaptability across benchmark NLP tasks, thanks to its being pre-trained on a large corpus

of texts, as well as its strong modelling capability brought about by its hundreds of millions of trainable parameters. However, generally, BERT can only process texts with a maximum sequence length of 512 tokens. The maximum sequence length for our human-generated texts is 453 tokens, whereas, for ChatGPT-generated texts, it extends to 644 tokens. Given this, 4.39% of the ChatGPT-generated texts (i.e., texts longer than 512 tokens) will be truncated during BERT tokenisation. As our text data are in Chinese, we deploy the multilingual BERT, trained using word pieces from 100+ lexica, to derive textual information. On top of the BERT backbone, we add an MLP network for the downstream binary classification task and optimise the fine-tuned BERT model by selecting the best combination of learning rate and batch size in the search space of [4e-5, 3e-5, 2e-5, 1e-5, 1e-4] and [16, 32]. Further to this, for the combined BERT model, we concatenate the 768-dimensional pooled BERT representation to the structured data as input to the MLP layer, which we train together with BERT. An illustration of our combined BERT model architecture is shown in Figure 3.

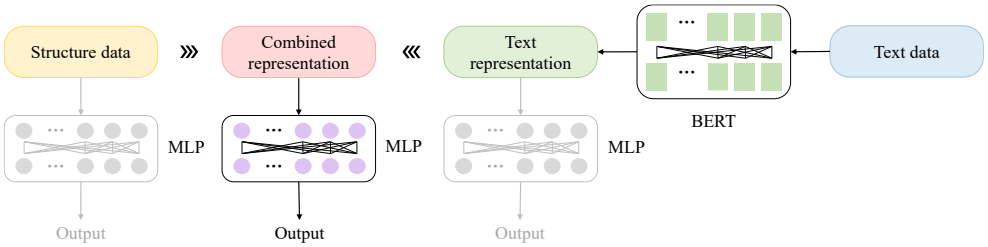


Figure 3 Combined BERT model architecture

4.2 Model training

We divide our dataset into three distinct subsets – training, validation and test – following the approach taken by Kriebel & Stitz (2022). The training set is used to train the models, the validation set is used to choose optimal model hyperparameters, and the test set is used to estimate out-of-sample prediction results. To construct these sets, we use the stratified random sampling method to select 70% of the samples as the training set, 20% of the training set as the validation set, and the remaining 30% as the test set. Thus, we have 1,377 observations in the training set, 345 in the validation set and 738 in the test set.

Given the nature of our binary-label classification task, we chose the binary cross entropy (BCE) loss as the primary loss function for our MLP models. The BCE loss for the entire set of observations is denoted by $L(y, \hat{y})$, where N represents the batch size, y_i is the actual label (0 for non-defaulters and 1 for defaulters), and $\hat{y}_i$ is the predicted probability of default for observation i. We run training for multiple epochs and select hyperparameters that minimise the BCE loss on the validation set. Models with these optimised hyperparameters are saved and utilised to predict the probabilities of default for the test set.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \quad (1)$$

4.3 Evaluation metrics

In this study, we consider two widely used metrics to evaluate the prediction results of our credit scoring models: the area under the receiver operating characteristic (ROC) curve (AUC) and the Kolmogorov-Smirnov (KS) statistic. The ROC curve is a graphical representation of a classification model's predictive accuracy over a range of threshold values. The curve plots the true positive rate (i.e., sensitivity) against the false positive rate (i.e. 1 – specificity) for each threshold value of a given model (Wu & Li, 2021). The higher the AUC score, the better the model distinguishes negative and positive classes. The KS statistic is defined by the maximum vertical distance between the empirical cumulative distribution functions of the false positive rate and the true positive rate and can be used to assess the correctness of categorical predictions (Lessmann et al., 2015; Wang et al., 2020). Similar to AUC, the KS statistic ranges from 0 to 1, with a larger KS value indicating stronger predictive capability. To reduce randomness and ensure the reliability and stability of the results, we adopt five random seeds to obtain average AUC and KS scores over all our experiments.

4.4 Model interpretability

In the business context, the interpretation and transparency of a model play an essential role in data-driven decision-making. However, a deep learning model is often considered to be a black-box and is not always well understood due to its complexity. In our study, to provide a better understanding of how textual input influences the prediction results of credit scoring models, we conduct content analysis using LIME. LIME, proposed by Ribeiro et al. (2016), is a model-agnostic technique that can be applied to any black-box model so that model predictions can be comprehended by perturbing the input data. It is implemented as a surrogate model at a local level, assuming a linear relationship between the model's input and output despite the non-linearity in the global relationship. It alters a single data sample by adjusting the feature values, and we can observe the effect on the output. When applied to the content analysis in this study, LIME produces the new text by randomly removing words/phrases from the original text input, with each word/phrase represented as a binary feature (1 for included and 0 for removed).[4]

[4] SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) is an alternative approach to exploring text importance. However, given that its usage would entail high computational costs for deep learning models and ignore the direction of the relationship between the input variables and the response (Stevenson et al., 2021), we have not utilised it in our study.

5. Empirical results

Figure 4 outlines our proposed research framework, showing how each research question is addressed. Details and results are discussed in Section 5.1 - 5.4 respectively for the four research questions. We also conduct further analysis in Section 5.5, introduced later.

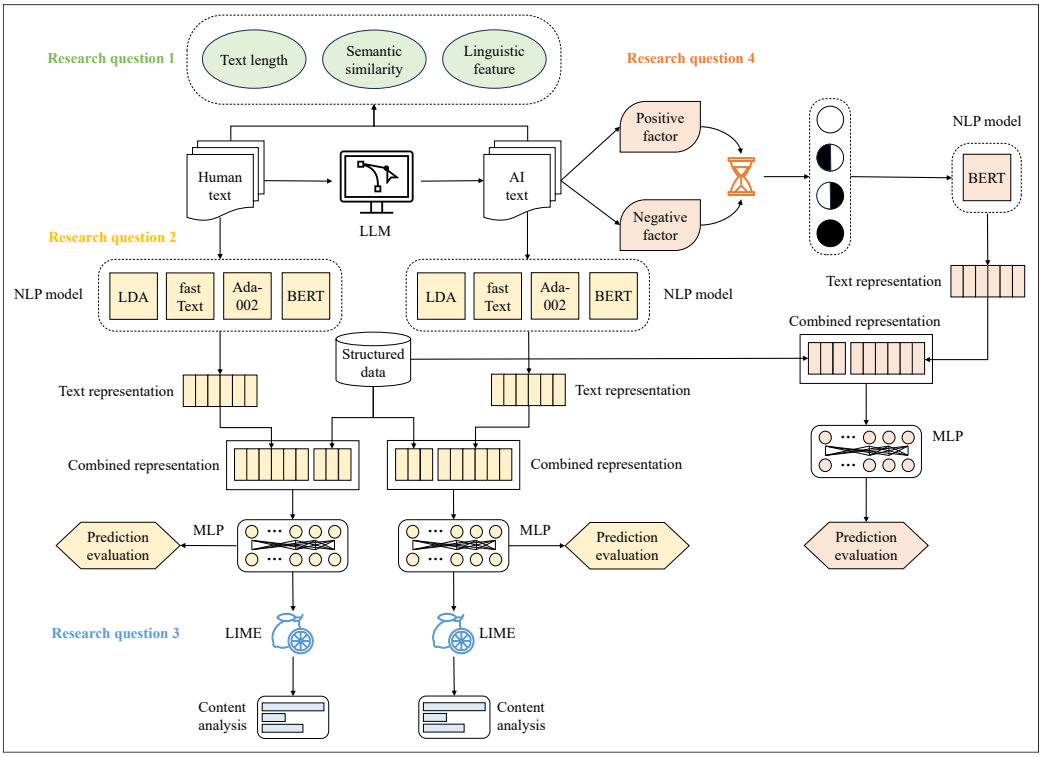


Figure 4 Our empirical framework for default prediction

5.1 RQ1: linguistic comparisons between human-generated and ChatGPT-generated texts

In this section, we investigate the differences between loan assessments generated by humans and ChatGPT in terms of text length, semantic similarity, and linguistic representations. Figure 5 illustrates that the average text length of ChatGPT-generated texts is significantly longer than that of the original texts, their values being 355 and 209, respectively.[5] To evaluate the difference in average text lengths between the two texts, we conduct the Mann-Whitney U test (Mann & Whitney, 1947). This test yields a statistic of 654,395 with a *p*-value very close to 0, suggesting a significant difference at the 1% level between the two average text lengths. This notable disparity emerges because ChatGPT-generated texts generally offer more comprehensive and thorough assessments of a borrower's creditworthiness, thereby enhancing the richness of information content.

[5] This result is also consistent with the findings of Guo et al. (2023), suggesting that ChatGPT tends to offer longer and more detailed responses.

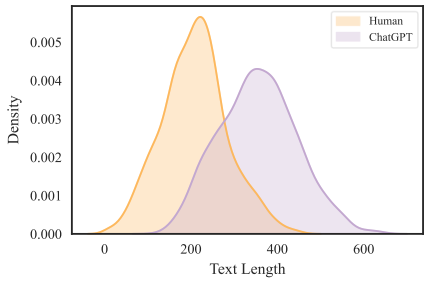


Figure 5 Text length of human-generated and ChatGPT-generated texts

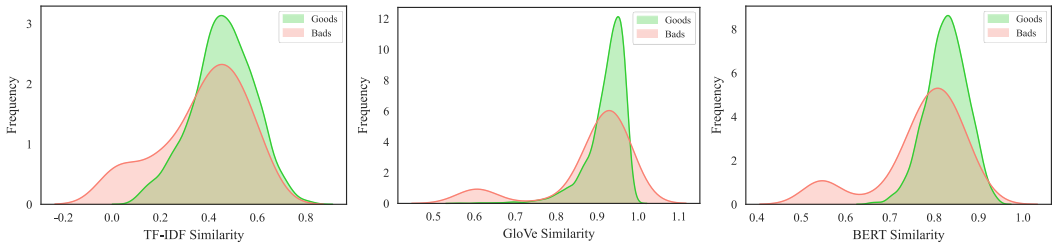
Furthermore, we calculate the cosine similarity between the embeddings of the two texts to evaluate the texts' semantic similarity. Cosine similarity is a vector-based measure used in a high-dimensional text representation space where similar texts are still close after projection (Tata & Patel, 2007). The cosine of the angle between two vectors reflects the semantic similarity of the corresponding texts. In this study, we employ three commonly used methods for text vectorisation: TF-IDF (Tata & Patel, 2007), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2018).[6] After obtaining the cosine similarity scores between human-generated and ChatGPT-generated texts, we employ the Mann-Whitney U test to further examine whether the average cosine similarity of the two texts for the Bads and Goods groups are the same. The results shown in Table 2 suggest a notable difference in the average text similarity across all text vectorisation methods. Additionally, it is noteworthy that the cosine similarity scores for the Bads group are consistently lower than those for the Goods group. Figure 6 visually confirms the above findings and shows that the Bads group exhibits a bimodal distribution while the Goods group's distribution is more likely left-skewed. These results suggest that ChatGPT may possess some abilities to presciently distinguish defaulters from non-defaulters which may analyse the original textual loan assessments with a different focus.

Table 2 Cosine similarity statistics across Bads and Goods groups

Model	Mean		SD		Statistic	*p*-value
	Bads	*Goods*	*Bads*	*Goods*		
TF-IDF	0.375	0.458	0.176	0.129	54,637	0.001***
GloVe	0.886	0.923	0.107	0.048	59,735	0.024**
BERT	0.772	0.828	0.096	0.046	46,120	0.000***

Notes: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

[6] TF-IDF is a method that uses word frequencies to quantify the importance of word representations in a document taken from a collection of documents (also known as a corpus). GloVe is a pre-trained predictive algorithm for forming word embeddings that leverages global word-to-word co-occurrence counts based on the entire corpus. Unlike TF-IDF and GloVe, which ignore the word's contextual meaning, transformer-based BERT learns sequence-level semantics by considering the sequences of all words in the document, enabling it to identify different representations for polysemous words.



(a) (b) (c)

Figure 6 Cosine similarity distributions: (a) TF-IDF, (b) GloVe, (c) BERT

To further investigate the linguistic representations presented in the loan assessments generated by humans and ChatGPT, we use the linguistic inquiry and word count (LIWC) dictionary to assess the texts' relevance across various linguistic and psychological dimensions (Li et al., 2018). The LIWC dictionary, initially proposed by Tausczik & Pennebaker (2010), has been extensively employed in various domains, such as computational linguistics and social relationships. As our text data are in Chinese, we employ the Chinese LIWC dictionary (Huang et al., 2012) to compute the frequencies of words that belong to different semantic categories for both texts. Then, we use Welch's t-test to examine any semantic differences between the Bads and Goods groups in the two types of texts.[7] Due to the space limit, the results of Welch's t-tests for each linguistic, psychological, or topical category of both texts for Bads and Goods are presented in Appendix Tables C.1 and C.2, respectively.

From Table C.1, we find that for the Bads group, there are significant differences in word frequencies across 36 out of 72 word categories when comparing the human- and ChatGPT-produced texts. The LIWC dictionary's division of word categories shows that the most notable difference between the two texts can be observed in words associated with cognitive words, pronouns, psychological words, and affective words. Specifically, ChatGPT-generated texts exhibit a higher usage of cognitive words, such as words in 'discrepancy', 'cognitive processes', 'tentative', 'differentiation', 'causation', and 'tendentiousness' categories. This indicates ChatGPT's ability to logically analyse human-generated texts and incorporate multiple perspectives. Furthermore, we observe that ChatGPT uses fewer first-person and informal pronouns, preferring more third-person pronouns than loan officers. This pattern

[7] Welch's t-statistic is calculated using the average values (f_1 and f_2) for a specific word category in the human and ChatGPT texts, respectively. The total number of text tokens contributed by the human and ChatGPT texts are denoted by N_1 and N_2. Following Li et al. (2018), we use the Bonferroni correction (M = 72) to adjust the significance level for each comparison. A positive t-statistic indicates an increased use of a semantic category in the ChatGPT text compared to the original human text. In contrast, a negative t-statistic indicates a decreased use of a semantic category in the ChatGPT text compared to the original human text. Welch's t-statistic is given as $t = \frac{f_2 - f_1}{\sqrt{\frac{f_1(1-f_1)}{N_1}} + \sqrt{\frac{f_2(1-f_2)}{N_2}}}$.

may reflect ChatGPT's objective standpoint. Regarding psychological words, we find an increased usage of categories like 'drives', 'achievement', and 'power' in ChatGPT's output, while 'affiliation' words show a decreased usage. These findings suggest ChatGPT's proficiency in identifying the borrower's personality traits and the potential for non-default outcomes. Furthermore, we find ChatGPT tends to use more affective words, such as 'negative emotion', 'anxiety', 'negations', and 'affect' words than the loan officers. Additionally, we observe notable differences in conversational words, perception words, and time words between human-generated and ChatGPT-generated texts.

Table C.2 shows that the usage of 64 word categories significantly differs between both texts for the Goods, which indicate that the semantic similarity between these two types of texts is relatively lower for non-defaulters than for defaulters. The differences are primarily observed in categories associated with cognitive, affective, pronoun, and perceptive words. Consistent with the results presented in Table C.1, ChatGPT-generated texts exhibit an increased usage of cognitive words, including 'discrepancy', 'cognitive processes', 'tentative', 'causation', 'differentiation', 'tendentiousness', 'certainty', 'insight', and 'comparison' words. Moreover, ChatGPT-generated texts tend to include more affective words like 'negative emotion', 'anxiety', 'affect', 'negations', 'sadness', 'anger', as well as 'positive emotion' words than in the human-generated texts. Conversely, words related to perception such as 'see', 'feeling', and 'space' words are more frequently used in human-generated texts, while 'motion' words are more commonly used in the ChatGPT-generated texts. Other categories, including function, physical, and social words, also exhibit significant differences between human-generated and ChatGPT-generated texts.

5.2 RQ2: prediction results of human-generated and ChatGPT-generated texts

In this section, we seek to examine the efficacy of text data for predicting default and compare the prediction results of human-generated and ChatGPT-generated texts. Table 3 presents the performance of the MLP models with the structured-only features, text-only features, and their combination, respectively.[8]

Table 3 Model performance of human-generated and ChatGPT-generated texts

Model	Category	Structured	Text	Combined	Structured	Text	Combined
			AUC			KS	
LDA+MLP	Human	0.615	0.608	0.617	0.251	0.091	0.288
	ChatGPT	0.615	0.621	0.620	0.251	0.166	0.254
fastText+MLP	Human	0.615	0.613	0.669	0.251	0.171	0.311

[8] We also examined the predictive power of the combination of human-generated and ChatGPT-generated texts in predicting defaults. The unreported results show that the model performance using the combined text is similar to that using ChatGPT-generated text only and further suggest that the information embedded within the ChatGPT-generated text inherently encompasses the content of the human-generated text.

	ChatGPT	0.615	0.611	0.635	0.251	0.213	0.278
Ada-002+MLP	Human	0.615	0.624	0.694	0.251	0.310	0.333
	ChatGPT	0.615	**0.683**	0.681	0.251	**0.346**	0.339
BERT+MLP	Human	0.615	0.630	0.667	0.251	0.283	0.361
	ChatGPT	0.615	0.662	**0.711**	0.251	0.331	**0.418**

Notes: This table presents the average AUC and KS scores for predicting the default risk of borrowers based on four NLP approaches: LDA+MLP, fastText+MLP, Ada-002+MLP, and BERT+MLP. Seventeen models are trained on different subsets, as illustrated in Table 1.

The average AUC and KS scores for models that rely solely on structured data are 0.615 and 0.251, respectively. As for models trained on the text-only features, the majority of the prediction results of using LDA+MLP and fastText+MLP models are worse than those generated by the structured-only models. In contrast, the text-only Ada-002+MLP and BERT+MLP models notably outperform the structured-only models, with Ada-002+MLP achieving the best average AUC and KS scores. These results highlight the success of NLP techniques and suggest that contextualised models (e.g., Ada-002 and BERT) are more effective in capturing textual information than bag-of-words (LDA) and static word embedding (fastText) approaches. When we examine the performance of the combined models, as expected, we find that almost all combined models outperform both structured-only and text-only models and achieve impressive AUC and KS scores, particularly for the BERT+MLP models. These findings suggest that incorporating unstructured text data into the credit scoring model can significantly improve prediction performances.

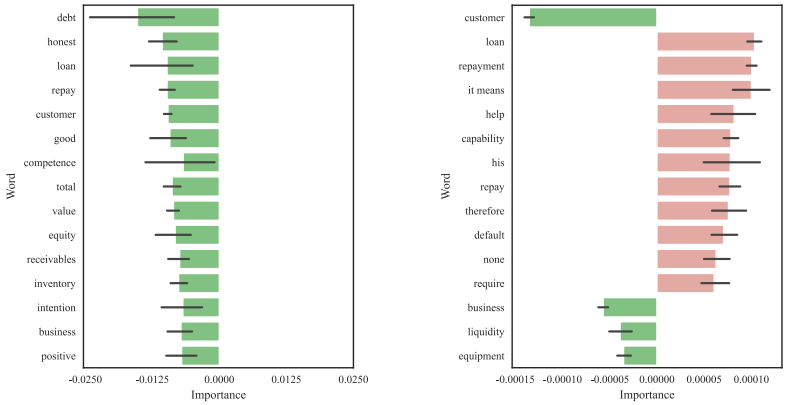
In Section 5.1, we show that raw textual loan assessments and ChatGPT-generated assessments exhibit certain differences. Here, we further compare the models' predictive power using these two types of texts to identify default risk. Table 3 shows that, in most cases, models which use ChatGPT-generated texts exhibit far stronger results than those using human-generated texts. Take the combined BERT+MLP model as an example: the average AUC and KS scores increase by 0.044 and 0.057, respectively, when using ChatGPT-generated texts as opposed to human-generated texts. These findings suggest that, compared with human-generated assessments, ChatGPT-generated texts may contain richer information and insights. As robustness checks, we further use four alternative ML classifiers, namely LR, DT, XGB and SVM, to supplement our main experiments; the related results supporting our main findings are available in Appendix D.

5.3 RQ3: text content analysis

In Section 5.2, our results indicate that text features extracted from loan assessments can serve as a powerful complement to traditional structured data, and its inclusion can significantly enhance the prediction results when attempting to identify the probability of default. In this section, we further explore which specific words/phrases can significantly influence the performance of credit scoring models.

Given that the combined BERT+MLP model outperforms all others, the following analysis will focus on this model specification to elucidate the influence of text content on model performance. Following Stevenson et al. (2021), we focus on uncertain cases from the test set, and these instances are identified as cases where the structured-only model yields predicted probabilities between 0.40 and 0.60, while an improvement of predicted probabilities is observed when using the combined BERT+MLP model. Of this subset, we review 250 cases where we can see the greatest improvement in the combined model for our analysis. This subsample is ideal for determining which textual content truly offers enhanced insights for default prediction.

In this analysis, we first employ LIME to identify which specific words contribute most to the model's default prediction.[9] Figures 7(a) and 7(b) present the top fifteen words derived from human-generated and ChatGPT-generated texts, respectively, in terms of their influence on the predicted probabilities of default. The length of the bar represents the mean influence of the word on the model's prediction. Bars with positive values (in red) indicate that the inclusion of the word leads to an increase in the predicted probability of default, while bars with negative values (in green) denote a reduced predicted probability of default.



(a) (b)

Figure 7 Top 15 important words for: (a) human-generated text, (b) ChatGPT-generated text

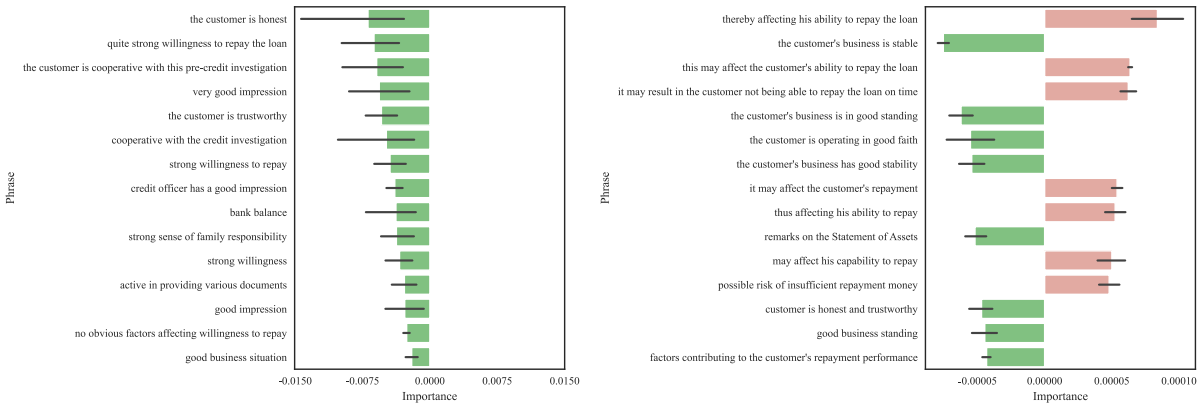
For those words sourced from human-generated texts, we find that all of the top fifteen words shown in Figure 7(a) contribute to a reduction in the predicted probability of default. This suggests that these words are associated with a heightened perception of creditworthiness. Several themes emerge, including business operations, loan requirements, borrower characteristics, and words with a positive

[9] The use of LIME offers managerial insights by identifying significant words and phrases that assist lenders in shaping future loan officer text descriptions. Stevenson et al. (2021) highlighted the potential of LIME to 'structure the unstructured data' which can enable lenders to capture more granular personal characteristics of borrowers than with standard model features. The analysis of text content can, therefore, provide a means to enhance loan assessments and offer valuable insights into the factors driving the credit risk assessment process.

semantic orientation. The prevalence of certain words, such as ‘customer’, ‘competence’, and ‘business’ all suggest that the model places significant emphasis on the borrower’s capacity to manage the business. This therefore implies that a stable operational status is correlated with a reduced predicted probability of default. The model also associates a low default risk with terms such as ‘equity’, ‘receivables’, and ‘inventory’, words that describe the borrower’s asset profile, which aligns with the common-sense view of loan review mechanisms. Furthermore, words pertaining to the borrower’s loan requirements themselves, including ‘debt’, ‘loan’, and ‘intention’, may reflect the borrower’s desire to obtain a loan for business development or to alleviate cash flow pressures and thus have a lower associated default prediction. The word ‘repay’ also reduces the predicted likelihood of default, indicating that repayment behaviour is an essential factor in the model’s overall assessment. Some other words describing the personality traits of the borrower are strongly associated with the model’s prediction of default risk. For instance, ‘honest’ and ‘good’ are explicit endorsements of the borrower and correspond to lower default prediction. The word ‘positive’ is another way the loan officers describe their assessment of the borrower’s attitude and can reflect the borrower’s willingness to provide documents and cooperate with the loan investigation process.

For the ChatGPT-generated texts, the fifteen most important words can be categorised into themes related to business operations, loan requirements and other words that lack ostensible factual meaning. In contrast to the results of the human-generated texts, Figure 7(b) shows that the majority of these fifteen words show positive importance values, suggesting they are associated with higher predicted probabilities of default. Words such as ‘customer’, ‘business’, ‘equipment’, and ‘liquidity’ indicate borrower’s stable operations and productions can lead to decreased default predictions. In contrast to the roles loan requirement terms play in human-generated texts, these terms indicate a higher predicted probability of default in almost all instances of ChatGPT-generated texts. Loan requirement words such as ‘loan’, ‘require’, and ‘help’ convey the borrower’s eagerness to relieve financial pressure by obtaining a loan and thus suggest a higher predicted probability of default. Some other loan repayment words such as ‘repayment’, ‘default’, and ‘none’ also lead to higher default risk. Moreover, some words that have no concrete meanings are nevertheless vital in affecting the model’s predictions, terms such as ‘it means’ and ‘his’. Although these words convey limited factual content, they notably affect the model’s output, indicating that these words may often relate to other words that have significant influence on the model’s predictions. Furthermore, the model associates a higher risk with the term ‘capability’, a finding which implies that, within the ChatGPT-generated texts, this term might frequently be employed to express scepticism regarding a borrower’s capability to repay.

Interestingly, the words ‘repay’ and ‘loan’ themselves are associated with positive values when they appear in the ChatGPT-generated texts, suggesting that their inclusion can increase the predicted probability of default, in contrast to the results derived from the human-generated tests. These observations indicate that the importance of words for default risk prediction may be intricately tied to their contextualised embeddings. To further explore this possibility, we employ LIME to examine the importance of phrases in both types of texts. Figures 8(a) and 8(b) present the top fifteen phrases that emerge as pivotal in both texts.



(a) (b)

Figure 8 Top 15 important phrases for: (a) human-generated text, (b) ChatGPT-generated text

Figure 8(a) shows that all of the fifteen most important phrases are associated with negative values, suggesting that their inclusion in human-generated text could reduce the model’s predicted probability of default. This finding is consistent with our prior findings concerning word importance, as illustrated in Figure 7(a). The most important phrases are mainly related to the loan officer’s comments on the borrower’s credibility, reliability, and willingness to repay the loan, as observed during the pre-loan investigation. To illustrate, phrases such as ‘strong willingness’ and ‘good impression’ repeatedly occur in Figure 8(a). It is noteworthy that several words listed in Figure 7(a), such as ‘customer’, ‘repay’, and ‘business’, also appear in Figure 8(a). The contexts in which these words are embedded suggest a lowered predicted probability of default.

In contrast to the results for the human-generated text, Figure 8(b) shows that only about half of the top fifteen phrases from the ChatGPT-generated text have negative importance values. In line with the findings from the previous section, most of the negative phrases are related to the description of the borrower’s business operation. For example, phrases such as ‘the customer’s business is stable’, ‘the customer’s business is in good standing’ and ‘good business standing’ all indicate a healthy business operation, which reduces the predicted probability of default. On the contrary, phrases that increase the predicted risk of default are primarily associated with the borrower’s compromised repayment capacity,

such as ‘thereby affecting his ability to repay the loan’ and ‘it may result in the customer not being able to repay the loan on time’. We also find that the words ‘loan’ and ‘repay’ are frequently present in phrases that convey risk, which is markedly different from their representations in human-generated texts. Such insights highlight the notion that the valence – positive or negative – of terms in default prediction is profoundly influenced by their contextualised setting.

5.4 RQ4: prediction results of positive and negative ChatGPT-generated factors

In this section, we further explore the influence of different components of the ChatGPT-generated text in enhancing the model’s prediction results based on our best-performing model BERT+MLP. As detailed in Section 3.3, we utilise ChatGPT to generate analytical assessments for human-generated texts from two perspectives: positive factors supporting the borrower’s repayment and negative factors that could potentially lead to the borrower’s default. In this section, we examine how varying combinations of these factors affect model predictions. Table 4 presents the discrimination performance for five types of texts (i.e., human, positive factors only, negative factors only, positive followed by negative factors and negative followed by positive factors) [10] using average AUC and KS scores.

Table 4 Model performance of different combinations of ChatGPT-generated factors

Category	Structured	Text	Combined	Structured	Text	Combined
		AUC			KS	
Human	0.615	0.630	0.667	0.251	0.283	0.361
Positive	0.615	0.611	0.663	0.251	0.263	0.347
Negative	0.615	0.669	0.707	0.251	**0.357**	0.407
Positive+Negative	0.615	0.662	0.711	0.251	0.331	**0.418**
Negative+Positive	0.615	**0.685**	**0.720**	0.251	0.347	0.408

Notes: This table shows the average AUC and KS scores of BERT+MLP for predicting the default risk of borrowers using original human-generated text and different combinations of ChatGPT-generated factors.

Table 4 shows that, except for the case of positive factors only, all of the text-only models surpass the structured-only model. In alignment with our earlier findings, the combined models consistently exhibit stronger performance than both the structured-only and text-only models. Further analysis of different text types shows that models which utilise the negative+positive text generally outperform those using other text choices, particularly in terms of average AUC scores. What is more, we observe that models focusing on the negative factor and the positive+negative text exhibit comparable performance on both text-only and combined subsets. Another noteworthy finding is that, once models incorporate the negative factor, they display stronger performance than those using the human-generated text. In light of these findings, we can conclude that incorporating comprehensive information derived

[10] As previously discussed, BERT truncates texts exceeding 512 tokens. This truncation could lead to a potential loss of information essential for default prediction when handling texts longer than 512 tokens. Therefore, the positive+negative and negative+positive texts can generate different prediction outcomes.

from both positive and negative factors can produce superior results, with the negative factor identified by ChatGPT playing a crucial role in enhancing prediction performance. Lenders should therefore prioritize the understanding and evaluating of negative elements that might impact a borrower's capability to fulfil payment obligations when it comes to gathering, processing, and analysing information.

5.5 Further analysis: prediction results of different LLM-generated texts

In this section, we conduct several additional experiments to validate our main findings illustrated in Section 5.1. In our main analysis, we rely on ChatGPT (version 2.13) to process and analyse the human-generated loan assessments. We now further investigate whether our main findings are sensitive to alternative generative AI models/versions, including ChatGPT4[11], ChatGLM2[12], and Baichuan2[13]. Table 5 presents the prediction results based on different LLM-generated texts.

Table 5 Model performance of different LLM-generated texts

Model	Category	Structure	Text	Combined	Structure	Text	Combined
			AUC			KS	
LDA+MLP	Human	0.615	0.608	0.617	0.251	0.091	0.288
	ChatGPT4	0.615	0.625	0.650	0.251	0.085	0.313
	ChatGLM2	0.615	0.591	0.628	0.251	0.203	0.262
	Baichuan2	0.615	0.634	0.653	0.251	0.162	0.297
fastText+MLP	Human	0.615	0.613	0.669	0.251	0.171	0.311
	ChatGPT4	0.615	0.567	0.634	0.251	0.096	0.263
	ChatGLM2	0.615	0.545	0.640	0.251	0.074	0.296
	Baichuan2	0.615	0.628	0.617	0.251	0.034	0.236
Ada-002+MLP	Human	0.615	0.624	0.694	0.251	0.310	0.333
	ChatGPT4	0.615	**0.706**	**0.734**	0.251	0.366	0.399
	ChatGLM2	0.615	0.592	0.690	0.251	0.227	0.326
	Baichuan2	0.615	0.573	0.672	0.251	0.208	0.284
BERT+MLP	Human	0.615	0.630	0.667	0.251	0.283	0.361
	ChatGPT4	0.615	0.693	0.719	0.251	**0.386**	**0.434**
	ChatGLM2	0.615	0.651	0.633	0.251	0.281	0.311
	Baichuan2	0.615	0.610	0.633	0.251	0.252	0.312

Notes: This table presents the average AUC and KS scores for predicting the default risk of borrowers based on four NLP approaches: LDA+MLP, fastText+MLP, Ada-002+MLP, and BERT+MLP across four types of texts, three of which are generated by LLMs.

In the table, we first observe that for most instances, combined models consistently show stronger

[11] ChatGPT4, building on OpenAI's GPT-4 model, has more parameters and training data than its predecessor, ChatGPT. This allows ChatGPT4 to better comprehend context, discern nuances and provide more accurate and coherent responses. The version of ChatGPT4 used in this paper is version 8.3, based on the GPT-4 model updated on 3rd August 2023.

[12] ChatGLM2 is the second-generation version of the open-source bilingual (Chinese-English) chat model ChatGLM proposed by Zhipu AI in 2022. It retains the smooth conversation flow and low deployment threshold of ChatGLM, while introducing stronger performance, longer context, more efficient inference and a more open license. More detailed information about ChatGLM2 can be found at https://huggingface.co/THUDM/chatglm2-6b.

[13] Baichuan2 is the new generation of large-scale open-source language models launched by Baichuan Intelligence inc. It is trained on a high-quality corpus with 2.6 trillion tokens and has achieved the best performance in authoritative Chinese and English benchmarks of the same size. More detailed information about Baichuan2 can be found at https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat.

performance than both structured-only and text-only results. For example, the average AUC and KS scores increase by 0.104 and 0.183, respectively, for the combined BERT+MLP model built on the ChatGPT4-processed text compared to its structured-only result. Moreover, in most instances, the models using generative AI-refined texts exhibit stronger outcomes than those using the original human-generated texts. These results are consistent with our previous findings, that incorporating textual information into credit scoring models could lead to superior performance, while AI-generated texts contribute to a more significant improvement for default prediction results.

When we explore the capabilities of various generative AI models in analysing texts and the effectiveness of their generated texts in predicting default, we find some interesting results, as shown in Table 5. Overall, ChatGPT4-generated texts generally exhibit superior performance across various NLP techniques and model subsets, particularly with Ada-002+MLP and BERT+MLP models. The performance of ChatGPT4-generated text is comparable with that of ChatGPT, which aligns with our expectations. Both ChatGPT and ChatGPT4 were developed by OpenAI and given their similar modelling infrastructures, it is reasonable to expect a high degree of similarity in the content they generate. When comparing the prediction results of ChatGLM2- and Baichuan2-generated texts, we find that their results vary depending on which NLP method is used. For example, for the LDA+MLP model, the Baichuan2 model achieves higher average AUC scores for both the text-only and combined subsets than the ChatGLM2 model. Conversely, Ada-002+MLP and BERT+MLP models show superior performance when using ChatGLM2-generated text. Moreover, models relying on ChatGLM2- and Baichuan2-generated texts do not demonstrate notable advantages over those based on human-generated texts for predicting default across all NLP approaches. In summary, different generative AI models can differ significantly in their understanding, analysis and response to a given text. Compared with other generative AI models, GPT-based models exhibit the most impressive performance in identifying borrowers' risks, thus enhancing overall default prediction results.

6. Conclusions

The rise of generative AI and advanced NLP techniques offers powerful new tools for extracting valuable insights from textual data, which can be instrumental in enhancing financial decision-making. In this study, we introduce a novel LLM-assisted approach to examine and assess the efficacy of textual information and the use of generative AI tools in predicting default for mSE borrowers. To the best of our knowledge, this study is the first application of generative AI to support lending decision-making and thus enhance credit risk mitigation.

In our empirical experiments, we first use ChatGPT to assess a (human) loan officer's produced

assessments and thereby generate AI-refined loan assessments. Our findings reveal that there are significant differences between human-generated and ChatGPT-generated texts in terms of text length, semantic similarity, and linguistic representations. Besides, when integrating unstructured text data into credit scoring models, we find that although both human-generated and AI-generated texts can enhance models' prediction results, the latter proves to be more effective, suggesting that AI-generated text contains amplified information which is valuable for loan default prediction. Moreover, we explore what kinds of words/phrases significantly contribute to the predicted probability of default. The results show us that words related to business operations, loan requirements, and borrower characteristics provide more insights into a given model's predictions, and the importance of such words to these predictions is influenced significantly by their contextual information. Finally, we examine which component of AI-generated texts has the most substantial influence on the model's performance enhancement. Our findings reveal that ChatGPT's analysis of borrower delinquency factors offers the most significant contribution to the model's prediction results. Moreover, we use several alternative generative AI tools for comparative analysis. These results further reinforce our conclusions and highlight the efficacy of GPT-based models in analysing textual information.

Our research not only contributes to the growing body of literature on the application of generative AI and NLP tools in the finance and business domain, but also has far-reaching managerial implications. First, as emerging LLMs are bringing about unprecedented changes over a wide range of industries (Chen et al., 2023), this work highlights the importance of utilising LLMs and offers a new hands-on approach to using them in the finance and business discipline. Following the concepts of software-as-a-service, platform-as-a-service, and infrastructure-as-a-service, we believe that language-model-as-a-service will be the next core concept that can help enhance human exploration and understanding of the world around us (Zhao et al., 2023). Second, this study demonstrates that using generative AI tools to summarise and analyse human-generated text can effectively enhance lending decision-making, illustrating the strong capabilities of LLMs in both language comprehension and generation. Besides, the LLM-assisted problem-solving procedure we have introduced can easily be adopted by financial institutions for investment decision-making and risk management analysis. Third, as alternative data such as images, video, audio and text become increasingly prevalent and more conveniently accessible (Chen et al., 2023), our work serves as a valuable practical application of how adequately leveraging text data, especially AI-generated texts, to support finance and business-related decision-making procedures. These perspectives are of great significance to both research and practice.

This work inevitably has limitations that can be addressed in future studies. First, the content

generated by LLMs can vary considerably due to the great flexibility of the prompts that can be used to guide them to generate output (Yang & Menczer, 2023). Future research could specifically explore the different results that would be obtained with different prompt templates for the same research topic. Second, as OpenAI's ChatGPT and ChatGPT4 are not yet open-source, we faced challenges in conducting completely reproducible experiments, given that version updates may alter the results going forward. Moreover, the generative AI tools we used are not tailored for the financial sector and analysis. Future research could explore open-source LLMs and/or finance specialised LLMs for default prediction. Third, the dataset size in our study is relatively small. Subsequent studies might benefit from employing larger datasets.

Declarations of interest

None.

References

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*(1), 168–178.

Alshurafat, H. (2023). The usefulness and challenges of chatbots for accounting professionals: Application on ChatGPT. *SSRN Electronic Journal*.

Anderson, R. A. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Bazley, W. J., Cronqvist, H., & Mormann, M. (2021). Visual finance: The pervasive effects of red on investor behavior. *Management Science*, *67*(9), 5616–5641. ·

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Cao, Y., & Zhai, J. (2023). Bridging the gap – the impact of ChatGPT on financial research. *Journal of Chinese Economic and Business Studies*, *21*(2), 177–191.

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: The growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*, 1–26.

Chen, G., Huang, L., Xiao, S., Zhang, C., & Zhao, H. (2023). Attending to customer attention: A novel deep learning method for leveraging multimodal online reviews to enhance sales prediction. *Information Systems Research*, isre.2021.0292.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, *28*(1), 1–32.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* (arXiv preprint arXiv: 1810.04805). arXiv.

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama Conjecture. *Finance Research Letters*, *53*, 103662.

Fitzpatrick, T., & Mues, C. (2021). How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research*, *294*(2), 711–722.

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection* (arXiv:2301.07597). arXiv.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, *48*(SPL), S138–S149.

Huang, C.-L., Chung, C. K., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B. C. P., Chen, W.-C., Bond, M. H., & Pennebaker, J. W. (2012). The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, *54*(2), 185–201.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, *62*(6), 1554–1577.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*(1–2), 511–529.

Jiang, C., Wang, Z., & Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, *277*(1), 20–31.

Kim, A. G., Muhn, M., & Nikolaev, V. V. (2023). Bloated disclosures: Can ChatGPT help investors process information? *SSRN Electronic Journal*.

Ko, H., & Lee, J. (2023). Can ChatGPT improve investment decision? From a portfolio management perspective. *SSRN Electronic Journal*.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309–323.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136.

Li, Y., Mihalcea, R., & Wilson, S. R. (2018). Text-based detection and understanding of changes in mental health. In S. Staab, O. Koltsova, & D. I. Ignatov (Eds.), *Social Informatics* (Vol. 11186, pp. 176–188). Springer International Publishing.

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems with Applications*, *74*, 105–114.

Liu, Y., Li, X., & Zheng, Z. (Eric). (2023). Smart natural disaster relief: Assisting victims with artificial intelligence in lending. *Information Systems Research*, isre.2023.1230.

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *SSRN: 4412788*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743–758.

Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research*, *261*(2), 656–665.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, *132*, 199–208.

Mukherjee, A., Panayotov, G., & Shon, J. (2021). Eye in the sky: Private satellites and government macro data. *Journal of Financial Economics*, *141*(1), 234–254.

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, *56*(6), 960–980.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on EMNLP*, 1532–1543.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, 1135–1144.

Song, Y., Wang, Y., Ye, X., Zaretzki, R., & Liu, C. (2023). Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *Information Sciences*, *629*, 599–617.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758–771.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research*, *34*(2), 399–408.

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of TF-IDF based cosine similarity predicates. *SIGMOD Record*, *36*(2), 7–12.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 5998–6008.

Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems*, *37*(1), 282–308.

Wenzlaff, K., & Spaeth, S. (2022). Smarter than humans? Validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance. *SSRN Electronic Journal*.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A large language model for finance* (arXiv:2303.17564). arXiv.

Wu, Z., & Li, Z. (2021). Customer churn prediction for commercial banks using customer-value-weighted machine learning models. *The Journal of Credit Risk*, *17*(4), 15–42.

Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020). Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, *39*(2), 260–280.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241.

Yang, K.-C., & Menczer, F. (2023). *Large language models can rate news outlet credibility* (arXiv:2304.00228). arXiv.

Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the power of voice for financial risk prediction: A theory-driven deep learning design approach. *MIS Quarterly*, *47*(1), 63–96.

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit

worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274–13283.

Yue, T., Au, D., Au, C. C., & Iu, K. Y. (2023). Democratizing financial knowledge with ChatGPT by OpenAI: Unleashing the power of technology. *SSRN Electronic Journal*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A survey of large language models* (arXiv:2303.18223). arXiv.

Supplementary materials

Appendix A. Overview of structured features

Table A.1 Definitions and statistical information of structured features

Feature	Definition	Summary statistics			
Continuous		*Mean*	*SD*	*Min*	*Max*
Loan amount ♣	Loan amount (¥)	90,550.81	22,839.84	1,000.00	200,000.00
Loan rate ♣	Loan interest rates (%)	15.48	0.46	3.00	15.84
Loan term ♣	Loan term (months)	3.07	0.47	1.00	8.00
Rest amount	Remaining loan amount outstanding (¥)	1,926.88	12,881.14	0.00	100,000.00
Rest interest	Remaining loan interest outstanding (¥)	88.17	734.32	0.00	11,322.52
Rate-of-income	The ratio of monthly repayment to income (%)	16.18	23.11	0.00	99.59
Firm age ♣	Number of years in operation	5.50	4.30	0.00	36.00
Space ♣	Office space (m^2)	1,025.67	3,094.97	0.00	33,300.00
Employees♣	Number of employees	11.50	26.56	0.00	500.00
Procurement	Annual procurement frequency	3.75	6.75	0.00	24.00
Family members	Number of borrower's family members	3.31	1.00	0.00	9.00
Family workforce	Number of workforces available in borrower's family	2.24	0.80	0.00	6.00
Owner age ♣	The age of the business owner	39.40	8.23	19.00	64.00
Categorical		*Number of categories*		*Values*	
Loan type	Loan category	4		{1, 2, 3, 4}	
Loan purpose ♣	Types of loan purpose	8		{1, 2, 3, 4, 5, 6, 7, 8}	
Repay type	Types of loan repayment	3		{1, 2, 3}	
Advance payment	Types of early loan repayment	2		{0, 1}	
Business type ♣	Types of borrowers' business	4		{A, B, C, D}	
Industry ♣	Industry of the borrower's business	12		{A, D, E, F, G, H, …}*	
Customer type	Whether the borrower is a new or existing customer of the bank	2		{new_cust, old_cust}	
Licence type	Types of business licence	3		{0, 1, 2}	
Credit rating	Borrower's credit rating	8		{1, 2, 3, 4, 5, 6, 7, 8}	
Ethnicity	Borrower's ethnicity	10		{1, 2, 3, 4, 6, 7, …}*	
Education ♣	Borrower's educational background	9		{10, 20, 30, 40, …}*	
Degree ♣	Borrower's academic degree	4		{0, 1, 2, 3}	
Marital status ♣	Borrower's marital status	7		{1, 2, 3, 4, 5, 6, 7}	
Homeownership ♣	Borrower's homeownership	7		{1, 2, 3, 4, 5, 6, 7}	
Occupation ♣	Borrower's occupation	8		{1, 2, 3, 4, 5, 6, Y, Z}	

Job position ♣	Borrower's job position	5	{A, B, C, D, E}
Job title ♣	Borrower's job title	5	{0, 1, 2, 3, 4}
Postcode	Postcode of the business location	19	{200000, 214400, …}*
Verified ID ♣	Borrower's citizenship category	5	{1, 2, 3, 4, 5}

Notes: We conduct a data desensitization process on the original structured data to protect private and business-sensitive information. * indicates that some values of categorical variables are omitted for the sake of space. ♣ indicates the final features selected for our credit scoring models.

Appendix B. A typical extract for non-defaulters' human-generated and ChatGPT-generated loan assessments

Human-generated text for a non-defaulter: *"The borrower is honest and has a clear business strategy. He was very cooperative with this loan investigation and actively provided various operational data. Historical data reflects the stability of his business operation. The borrower has been engaged in the metal building materials business for many years and has a wide range of connections. His main customers are well-acquainted friends. The borrower also has a strong willingness to repay. From our past investigations and feedback from the managers of the marketplace, the borrower's business is stable, and he has no bad habits. The credit system shows that the borrower has a good credit history and no overdue payment records. The borrower's previous loan payments were made on time, and he will renew the loan sooner. The balance of the borrower's credit card is ¥ ???. According to the unsettled delivery notes, the receivables are estimated to be ¥ ???, which basically matches the borrower's stated receivables of ¥ ???. The majority of the receivables will be settled within a month. We are conservatively estimating them to be ¥ ???. Additionally, the borrower has ¥ ??? of security deposit at a guarantee company. As physical inventory, the borrower has ??? tons of stock at a cost price of ¥ ??? and the estimated stock value is ¥ ???. We are cautiously recording this as an asset value of ¥ ???."*

ChatGPT-generated text for a non-defaulter: *"1. Factors supporting the borrower's repayment: * The borrower is trustworthy and has a clear business strategy. Historical information reflects his stable business operation and strong willingness to repay. * The borrower has been engaged in the metal building materials business for many years and has a wide range of connections. His main customers are well-acquainted friends, with whom he has a deep mutual understanding. * The credit system has revealed that the borrower has a good credit history with no record of overdue payments, and the borrower's previous repayments were very timely. 2. Factors that could lead to the borrower's default: * The balance on the borrower's bank card is ¥ ???. Financial strain might result in the borrower's inability to repay on time. * There are conservative estimates in the calculation of inventory value. If the actual value of the inventory is insufficient to cover the loan amount, it may also lead to the borrower's default."*

Appendix C. Welch's t-test results for word category differences

Table C.1 Welch's t-test results of human-generated and ChatGPT-generated texts for the Bads

Category	Human	ChatGPT	t-stat	*p*-value	Category	Human	ChatGPT	t-stat	*p*-value
Affective words					Religion	0.022	0.018	-2.449	1.132
Affect	0.065	0.082	5.573	0.000***	Work	0.079	0.092	3.921	0.007**
Anger	0.000	0.001	1.454	11.533	*Motive words*				
Anxiety	0.001	0.006	9.176	0.000***	Reward	0.019	0.021	1.367	13.558
Negations	0.010	0.021	7.779	0.000***	Risk	0.008	0.015	5.823	0.000***
Negative emotion	0.010	0.029	13.209	0.000***	*Number words*				
Positive emotion	0.042	0.040	-0.812	32.945	Numbers	0.012	0.008	-3.778	0.013*
Sadness	0.001	0.002	3.372	0.059	Quantities	0.019	0.023	2.801	0.403
Swear words	0.002	0.000	-2.904	0.292	Quantity units	0.121	0.097	-6.161	0.000***
Cognitive words					*Perception words*				
Causation	0.009	0.026	12.341	0.000***	Feeling	0.007	0.004	-3.185	0.114
Certainty	0.014	0.016	1.187	18.595	Hear	0.002	0.007	7.786	0.000***
Cognitive processes	0.082	0.182	26.878	0.000***	Motion	0.021	0.031	5.869	0.000***
Comparison	0.027	0.034	3.103	0.152	Perception	0.031	0.027	-2.350	1.482
Differentiation	0.007	0.025	12.997	0.000***	See	0.016	0.008	-5.769	0.000***
Discrepancy	0.017	0.080	29.196	0.000***	Space	0.055	0.051	-1.469	11.197
Insight	0.035	0.034	-0.310	59.759	*Physical words*				
Relativity	0.111	0.115	0.988	25.544	Body	0.009	0.009	-0.498	48.855
Tendentiousness	0.031	0.041	4.509	0.001***	Death	0.001	0.001	0.593	43.705
Tentative	0.007	0.030	16.658	0.000***	Food	0.007	0.008	1.050	23.194
Conversational words					Health	0.003	0.002	-0.707	37.89
Assent	0.024	0.037	6.411	0.000***	Physical	0.020	0.019	-0.413	53.672
Fillers	0.004	0.004	-0.021	77.658	Sexual	0.002	0.002	-0.614	42.591
Netspeak	0.017	0.025	4.998	0.000***	*Pronoun words*				
Nonfluencies	0.000	0.001	5.295	0.000***	1st person plural	0.001	0.000	-3.616	0.024*
Psychological words					1st person singular	0.003	0.000	-4.917	0.000***
Achievement	0.016	0.033	9.885	0.000***	3rd person plural	0.000	0.001	5.661	0.000***
Affiliation	0.012	0.006	-4.472	0.001***	3rd person singular	0.001	0.004	6.197	0.000***
Drives	0.095	0.132	10.013	0.000***	Impersonal pronouns	0.008	0.004	-4.600	0.000***
Power	0.046	0.075	10.796	0.000***	Personal pronouns	0.005	0.005	0.470	50.408
Function words					Pronouns	0.013	0.009	-3.209	0.105
Adverbs	0.068	0.080	3.624	0.023*	*Social words*				
Auxiliary verbs	0.043	0.102	20.987	0.000***	Family	0.001	0.000	-1.922	4.315

Conjunctions	0.020	0.051	15.747	0.000***	Female references	0.002	0.000	-2.952	0.25
Function words	0.383	0.495	19.214	0.000***	Friends	0.000	0.000	-0.170	68.343
Interrogatives	0.002	0.001	-2.073	3.013	Male references	0.000	0.000	-0.555	45.757
Postpositions	0.014	0.013	-0.916	28.416	Social processes	0.053	0.057	1.619	8.327
Prepositions	0.056	0.057	0.473	50.267	*Time words*				
Programmes	0.004	0.003	-1.597	8.711	Future focus	0.008	0.024	12.209	0.000***
Lifestyle words					Past focus	0.006	0.004	-1.961	3.944
Home	0.003	0.002	-0.995	25.249	Present focus	0.015	0.011	-2.637	0.662
Leisure	0.006	0.002	-4.948	0.000***	Time	0.048	0.036	-4.946	0.000***
Money	0.106	0.098	-2.115	2.722					

Notes: 'Human' and 'ChatGPT' denote the original human-generated and ChatGPT-generated textual loan assessment, respectively. *$p < 0.1$/M; **$p < 0.05$/M; ***$p < 0.01$/M.

Table C.2 Welch's t-test results of human-generated and ChatGPT-generated texts for the Goods

Category	Human	ChatGPT	t-stat	*p*-value	Category	Human	ChatGPT	t-stat	*p*-value
Affective words					Religion	0.023	0.023	0.055	75.555
Affect	0.063	0.084	47.346	0.000***	Work	0.079	0.092	25.630	0.000***
Anger	0.001	0.001	8.325	0.000***	*Motive words*				
Anxiety	0.001	0.006	60.811	0.000***	Reward	0.016	0.022	24.255	0.000***
Negations	0.014	0.020	30.622	0.000***	Risk	0.007	0.014	42.353	0.000***
Negative emotion	0.010	0.027	72.095	0.000***	*Number words*				
Positive emotion	0.042	0.045	7.389	0.000***	Numbers	0.013	0.008	-28.525	0.000***
Sadness	0.000	0.002	21.181	0.000***	Quantities	0.019	0.023	15.153	0.000***
Swear words	0.001	0.001	-14.222	0.000***	Quantity units	0.128	0.099	-49.532	0.000***
Cognitive words					*Perception words*				
Causation	0.009	0.025	72.801	0.000***	Feeling	0.005	0.003	-14.205	0.000***
Certainty	0.013	0.016	12.559	0.000***	Hear	0.003	0.006	33.665	0.000***
Cognitive processes	0.080	0.167	157.364	0.000***	Motion	0.022	0.028	25.483	0.000***
Comparison	0.030	0.031	5.590	0.000***	Perception	0.027	0.027	-2.945	0.255
Differentiation	0.011	0.021	45.134	0.000***	See	0.015	0.009	-30.416	0.000***
Discrepancy	0.016	0.071	168.183	0.000***	Space	0.057	0.049	-19.073	0.000***
Insight	0.033	0.036	8.239	0.000***	*Physical words*				
Relativity	0.111	0.109	-4.478	0.001***	Body	0.009	0.008	-5.572	0.000***
Tendentiousness	0.026	0.039	41.740	0.000***	Death	0.001	0.001	0.547	46.174
Tentative	0.005	0.023	92.908	0.000***	Food	0.006	0.006	0.644	41.028
Conversational words					Health	0.003	0.002	-12.987	0.000***
Assent	0.025	0.038	43.343	0.000***	Physical	0.019	0.017	-8.785	0.000***
Fillers	0.004	0.004	0.696	38.407	Sexual	0.002	0.003	6.711	0.000***
Netspeak	0.020	0.023	12.323	0.000***	*Pronoun words*				
Nonfluencies	0.000	0.000	1.386	13.102	1st person plural	0.002	0.000	-27.811	0.000***
Psychological words					1st person singular	0.003	0.000	-35.243	0.000***
Achievement	0.015	0.031	61.731	0.000***	3rd person plural	0.000	0.000	9.735	0.000***
Affiliation	0.013	0.008	-28.067	0.000***	3rd person singular	0.001	0.002	20.139	0.000***
Drives	0.090	0.133	79.726	0.000***	Impersonal pronouns	0.008	0.004	-27.482	0.000***
Power	0.044	0.074	74.774	0.000***	Personal pronouns	0.005	0.002	-27.328	0.000***
Function words					Pronouns	0.013	0.006	-38.650	0.000***
Adverbs	0.067	0.078	25.170	0.000***	*Social words*				
Auxiliary verbs	0.043	0.098	128.758	0.000***	Family	0.002	0.001	-11.758	0.000***

Conjunctions	0.016	0.046	105.479	0.000***	Female references	0.002	0.001	-19.993	0.000***
Function words	0.380	0.475	108.784	0.000***	Friends	0.001	0.000	-10.778	0.000***
Interrogatives	0.002	0.002	-5.618	0.000***	Male references	0.000	0.000	-6.836	0.000***
Postpositions	0.014	0.011	-14.078	0.000***	Social processes	0.058	0.060	3.177	0.118
Prepositions	0.055	0.055	0.235	64.315		*Time words*			
Programmes	0.004	0.004	-5.441	0.000***	Future focus	0.007	0.023	78.707	0.000***
	Lifestyle words				Past focus	0.006	0.004	-12.476	0.000***
Home	0.004	0.003	-12.590	0.000***	Present focus	0.010	0.010	-3.519	0.034*
Leisure	0.005	0.002	-27.458	0.000***	Time	0.045	0.034	-29.912	0.000***
Money	0.104	0.093	-20.275	0.000***					

Notes: 'Human' and 'ChatGPT' denote the original human-generated and ChatGPT-generated textual loan assessment, respectively. *$p < 0.1$/M; **$p < 0.05$/M; ***$p < 0.01$/M.

Appendix D. Robustness check across alternative classifiers

Table D.1 Search space of hyperparameters for alternative classifiers

Classifier	Search space
LR	$C \in \{0.01, 0.1, 1\}$
DT	$max_depth \in \{2, 6, 8, 10\}$
	$min_samples_split \in \{2, 6, 8, 10\}$
	$min_samples_leaf \in \{2, 6, 8, 10\}$
XGB	$learning_rate \in \{0.01, 0.1\}$
	$max_depth \in \{2, 6, 8\}$
	$n_estimators \in \{30, 70, 90, 110\}$
	$scale_pos_weight \in \{1, 38\}$
SVM	$C \in \{0.1, 1, 2, 3, 4\}$
	$class_weight \in \{None, "balanced"\}$
	$kernel \in \{"rbf", "poly"\}$
	$gamma \in \{"scale", "auto"\}$

Notes: This table shows the considered hyperparameter spaces for the alternative classifiers used for robustness check of Section 5.1 in our study. The final hyperparameter combination is chosen based on validation sample performance.

Table D.2 Model performance of human-generated and ChatGPT-generated texts on alternative classifiers

			Human		ChatGPT			Human		ChatGPT	
NLP method	Classifier	Structure	Text	Combined	Text	Combined	Structure	Text	Combined	Text	Combined
		AUC					KS				
LDA	LR	0.664	0.592	0.667	0.716	0.711	0.292	0.186	0.288	0.406	0.392
	DT	0.589	0.603	0.617	0.659	0.628	0.211	0.236	0.228	0.288	0.260
	XGB	0.669	**0.700**	**0.739**	0.720	**0.747**	0.297	**0.446**	**0.494**	0.357	0.446
	SVM	0.670	0.669	0.651	0.726	0.687	0.296	0.281	0.296	0.390	0.306
fastText	LR	0.664	0.549	0.665	0.681	0.671	0.292	0.182	0.271	0.389	0.311
	DT	0.589	0.598	0.598	0.600	0.600	0.211	0.203	0.203	0.197	0.197
	XGB	0.669	0.623	0.649	0.610	0.716	0.297	0.333	0.321	0.215	0.400
	SVM	0.670	0.692	0.658	0.649	0.627	0.296	0.376	0.256	0.392	0.212
Ada-002	LR	0.664	0.646	0.675	**0.733**	0.670	0.292	0.354	0.299	**0.443**	0.294
	DT	0.589	0.629	0.625	0.657	0.657	0.211	0.261	0.253	0.296	0.296
	XGB	0.669	0.663	0.625	0.654	0.654	0.297	0.258	0.269	0.356	0.356
	SVM	0.670	0.593	0.650	0.661	0.648	0.296	0.240	0.003	0.397	0.003
BERT	LR	0.664	0.598	0.677	0.696	0.734	0.292	0.315	0.294	0.440	**0.456**
	DT	0.589	0.581	0.595	0.578	0.540	0.211	0.174	0.192	0.152	0.081
	XGB	0.669	0.510	0.584	0.717	0.720	0.297	0.221	0.231	0.378	0.425
	SVM	0.670	0.573	0.677	0.665	0.700	0.296	0.219	0.301	0.276	0.331

Notes: This table presents the average AUC and KS scores for predicting the default risk of borrowers based on four NLP approaches and four classifiers including LR, DT, XGB and SVM. For the BERT-combined models, classifiers are trained using the integration of 768-dimensional pooled representation produced by BERT with the structured data.

In the Table D.2, we observe that combined models generally display a superior performance to text-only models, and most models relying on ChatGPT-generated texts surpass those relying on human-generated texts. When comparing classifiers, we find that LR, XGB, and SVM outperform DT in most cases, and XGB does so especially. Notably, the LDA+XGB combined model achieves the highest average AUC scores of 0.739 and 0.747 for human-generated and ChatGPT-generated texts, respectively. When it comes to the effectiveness of textual feature extraction, the results of using the LDA model are comparable to those obtained using the fastText, Ada-002, and BERT models. This may be due to the fact that fastText, Ada-002, and BERT derive 300-, 1536-, and 768-dimensional features from texts, respectively, thus causing the classifiers to suffer from the curse of dimensionality. In summary, these results bolster our

confidence in the conclusions drawn from our primary analysis. Specifically, integrating unstructured text data can significantly enhance credit default predictions, and the use of ChatGPT-generated texts can further magnify the effectiveness of using text in credit default prediction.