Interpretable LLMs for Credit Risk: A Systematic Review and Taxonomy

Muhammed Golec[1,2], Maha AlabdulJalil[3]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Institute for Data Science and Artificial Intelligence, Boğaziçi University, Turkey
[3]College of Science Computer Science Department, Kuwait University, Kuwait

This manuscript remains under review.

Abstract

Large Language Models (LLM), which have developed in recent years, enable credit risk assessment through the analysis of financial texts such as analyst reports and corporate disclosures. This paper presents the first systematic review and taxonomy focusing on LLM-based approaches in credit risk estimation. We determined the basic model architectures by selecting 60 relevant papers published between 2020-2025 with the PRISMA research strategy. And we examined the data used for scenarios such as credit default prediction and risk analysis. Since the main focus of the paper is interpretability, we classify concepts such as explainability mechanisms, chain of thought prompts and natural language justifications for LLM-based credit models.

The taxonomy organizes the literature under four main headings: model architectures, data types, explainability mechanisms and application areas. Based on this analysis, we highlight the main future trends and research gaps for LLM-based credit scoring systems. This paper aims to be a reference paper for artificial intelligence and financial researchers.

1 Introduction

Credit risk assessment is one of the important components in financial decision making, such as making investment decisions and determining whether to grant credit to an individual or institution [1]. While traditional methods rely on structured data such as financial ratios and past payment information when assessing credit risk, in reality important information about the credit may not be available in an organized manner [2]. An example of this is that executive comments, economic news and analyst reports are generally in free text (unstructured).

Advanced LLM models such as GPT and FinBERT have great potential in financial applications with their high performance in extracting meaning from free text [3, 4]. They can produce interpretable outputs by making sense of complex financial data and thus facilitate decision-making processes [5]. With its high potential, this research area, LLM-Driven Credit Risk Assessment, where Natural Language Processing (NLP), financial analysis and Explainable Artificial Intelligence (XAI) intersect, has begun to attract the attention of researchers. However, the applications of LLMs in the field of credit risk analysis have not yet been systematically examined. The majority of studies in the literature superficially examine the applications of artificial intelligence in the financial sector or do not address the interpretability of LLMs in credit risk assessment in detail. Moreover, this particular research area requires a comprehensive taxonomy of LLM model types, data sources and explainable artificial intelligence (XAI) techniques.

1.1 Objectives and Contributions

This paper systematically reviews LLM-based approaches in credit risk assessment and presents a detailed taxonomy study. The main contributions of this paper are as follows:

- Published studies for LLM-based credit risk applications are systematically reviewed with the PRISMA methodology.

- A detailed taxonomy is presented by examining the current research from four main aspects: model architecture, data modality, explainability mechanism and application area.

- Trends, challenges and open research directions in the deployment of LLMs are analyzed.

- The first systematic review and taxonomy focusing on LLM-based approaches in credit risk estimation is presented, providing a reference for researchers and financial institutions in the field.

1.2 Paper Structure

Figure 1 shows the organization of the paper. Section 2 examines the current survey studies in the field related finance and LLM , examining their focus and limitations. Then, a comparison analysis is performed by comparing this paper and the literature. Section 3 explains the research questions and article collection strategy. Section 4 provides a taxonomy by classifying LLM-oriented credit risk assessment systems in four main aspects: model architecture, data modality, interpretability mechanism and application area. Section 5 synthesizes the main emphases obtained throughout the paper and highlights the literature gaps. Section 6 concludes the paper by discussing the implications for researchers and finance practitioners in the field.
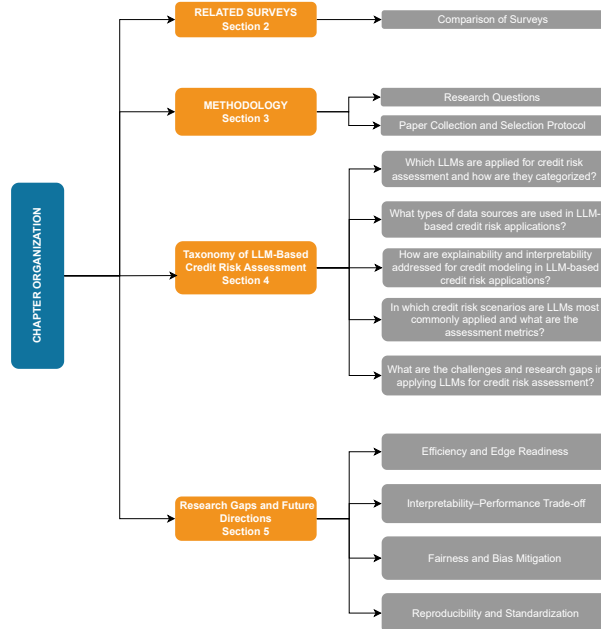

Figure 1: The Organization of the Paper

2   Related Surveys

Although there has been increasing interest in LLM and artificial intelligence (AI)-based financial studies in recent years, there is still no comprehensive classification in terms of focus, depth and relevance to credit risk, and discussions about interpretability, a key component in credit assessment scenarios, are lacking.

Staegemann et al. [6], in their literature review on generative AI applications in banking, highlight the potential of LLMs for risk reduction and customer experience. However, the article does not include techniques for credit risk modeling or interpretability, and there is no taxonomy of model types. In [7], models such as FinGPT and BloombergGPT are discussed, LLMs address financial inclusion and policy implications, and ethical concerns, but ignore credit risk applications. In Joshi et al. study's AI frameworks in credit risk and trading applications are examined, but does not include LLM architectures or datasets. [8]. Chen et al. [9], in their study of LLMs in finance, law, and healthcare, also highlight the risks of using LLMs in high-risk sectors. However, the financial discussion is superficial, and credit scoring methods are not addressed. In [10], sentiment analysis techniques used in banking are reviewed, and the focus is on examining the impact on investor confidence. However, credit risk tasks are only superficially examined. In Nie et al. study's the research on generative AI and LLMs focused on risk modeling and deployment architectures [11]. The paper discusses areas such as data engineering, credit scoring modules, etc. but lacks a comparative taxonomy and interpretability mechanisms for LLM types. Zavitsanos et al. [12] examines ML approaches in financial risk detection, examining features, data labels, evaluation techniques, and ML methods. Although this is a useful taxonomy, it is more focused on ML than LLM. In [13], the authors mainly emphasize traditional ML and statistical-based methods for financial forecasting, but do not address large language models or explainable AI methods. In Joshi et al. study's research on AI frameworks in credit risk and trading applications, but does not focus on LLM-specific taxonomy or performance benchmarking [14]. Kong et al. [15] presented a study that benchmarked LLM-based financial tasks for three different languages (Chinese, Japanese, English). However, the model types are not classified and also the explainability is not analyzed for credit risk domains. In Joshi et al. study's [16]. provides detailed architectural review on LLM-based applications for credit scoring and macroeconomic simulations. However, the paper lacks empirical comparison and also lacks a classification for interpretability. Lee et al. [17] review domain-specific financial applications based on LLM, but categorizes flow tasks and datasets, but does not focus on credit risk subdomains. In [18] investigates AI-based methods in financial scenarios such as fraud detection and portfolio management, but does not provide a detailed review of model interpretability and LLM for credit scoring. In Karami and Igbokwe study examines AI-based risk assessment and limitations of traditional credit methods, but does not provide a classification of LLM [19]. Alonso et al. [20] provides an overview of LLMs in areas such as asset management and risk reporting, but is insufficient for credit risk. In [21], LLMs are studied for financial applications such as credit analysis and fraud detection. However, they are very superficial in interpretability and credit-specific modeling. Lakkaraju et al. [22] evaluates LLM performance through a fairness lens, focuses on bias reduction strategies such as improving user trust, but credit risk assessments and explainability are very superficial. In [23], LLM-based financial sentiment analysis with multiple datasets is not considered credit risk or XAI. Omoseebi et al. [24] investigates the financial security side of LLMs such as fraud detection, but does not perform detailed credit risk modeling and taxonomy studies. In [25], democratization of financial datasets and open data models of LLMs are discussed, but no empirical comparison is included for credit or risk analytics. Krause et al. [26] present a survey discussing the sustainability of LLM models such as ChatGPT. The survey is mostly on ethical and governance concerns and is superficial in terms of credit scoring, model breakdowns, and datasets. In [27], the robustness of LLMs in financial tasks is assessed and benchmarks are made on summarization and event

detection. However, the study does not include credit-related tasks and explainability is weak.

Table 1: Comparison of Related Survey Studies with This Work

Work	Main Aim	Credit Risk	LLM-Specific	XAI	Taxonomy
[6]	GenAI in Banking	✗	✓	✗	✗
[7]	LLMs for Financial Regulation	✗	✓	✗	✗
[8]	Agentic GenAI for Risk Mgmt	✓	✓	✗	✗
[9]	LLMs in Finance/Healthcare/Law	✗	✓	✗	✗
[10]	Sentiment in Banking Headlines	✗	✓	✗	✗
[11]	LLMs in Financial Applications	✗	✗	✗	✗
[12]	ML for Financial Risk Reports	✓	✗	✗	✓
[13]	AI in Modern Banking	✓	✓	✗	✗
[14]	GenAI Agents in Finance	✓	✓	✗	✗
[15]	LLMs in Investment Mgmt	✗	✓	✗	✗
[16]	GenAI for Financial Risk	✓	✓	✗	✗
[17]	Survey of FinLLMs	✓	✓	✗	✓
[18]	Explainable AI in Finance	✓	✗	✓	✗
[19]	Big Data in Credit Risk	✓	✗	✗	✓
[20]	LLMs for Financial Reasoning	✗	✓	✗	✗
[21]	LLMs in Financial AI	✗	✓	✗	✓
[22]	LLMs as Finance Advisors	✗	✓	✗	✗
[23]	LLMs for Sentiment Analysis	✗	✓	✗	✗
[24]	LLMs for Financial Security	✗	✓	✗	✗
[25]	Consistency of LLMs	✗	✓	✗	✗
[26]	LLMs in Finance (ChatGPT/Bard)	✗	✓	✓	✗
[27]	LLM Strategy in FIs	✗	✓	✗	✗
This Work	Interpretable LLMs for Credit Risk	✓	✓	✓	✓

Comparison of Surveys: Table 1 provides a comparison of the surveys examined in this section. While interest in the financial domain applications of LLMs is increasing day by day, existing studies focus on broad themes such as generative AI [6], inclusiveness frameworks [7] and financial NLP [17, 21]. Only a limited number of studies directly focus on credit risk tasks, and none of them provide a taxonomy by examining LLM architectures, data formats, interpretability techniques and domain-specific applications.

Furthermore, most studies do not provide interpretability for high-risk decisions such as credit scoring, and the studies that do provide interpretability are very superficial [20, 25]. Some of the surveys include studies such as sentiment analysis or fraud detection [23, 10], but they do not address the concept of credit risk modeling. Some of the reviewed studies focus only on traditional machine learning (ML) [12, 19] and only propose conceptual frameworks without model comparison [8, 16].

Although some of the studies discuss domain-specific LLMs (such as FinBERT, FinGPT, and InvestLM), they do not systematically model architecture or risk. None of the existing studies provide a classification of LLM applications that addresses both modeling and regulatory requirements (e.g., explainability, data provenance, or fairness). **This study is the first to fill this gap based on four main pillars: model architecture, data modality, interpretability mechanism, and application domain.**

3 Methodology

This section describes the methodology used in the systematic review of LLM-based credit risk assessment.

3.1 Paper Collection and Selection Protocol

For the paper collection, refereed journals, high-ranking conferences, book chapters covering the years 2020-2025 were collected by scanning libraries such as IEEE Xplore, Elsevier, ACM Digital Library, SpringerLink, Scopus and arXiv. The following keywords were used while scanning the paper:

1. [(LLM) — (LargeLanguageModel)] & [(CreditScoring) —— (CreditRiskAssessment)]—

2. [(LLM) — (Transformer)] & [(Explainability) —— (XAI)] & [(FinancialNLP)]—

3. [(CreditRisk)] & [(MultimodalData) — (UnstructuredText) —— (BehavioralData)]—

4. [(LLM)] & [(GPT4 — FinBERT)] & [(Evaluation —— Benchmark)]—

5. [(Fairness) — (Hallucination) —— (Reproducibility)] & [(LLM) —— (CreditRisk)]—

Figure 2 summarizes the strategy followed for paper collection in this survey. Following the PRISMA guidelines, 182 papers were first collected as the initial body as a result of the relevant keys. After removing duplicates and papers that were not related to the research area, 120 papers were obtained. With the joint efforts of both authors in this survey (according to the inclusion criteria), 51 articles were obtained for the final analysis and this number was increased to 60 papers with the snowball method. A classification of all these papers is shown in Table 2.
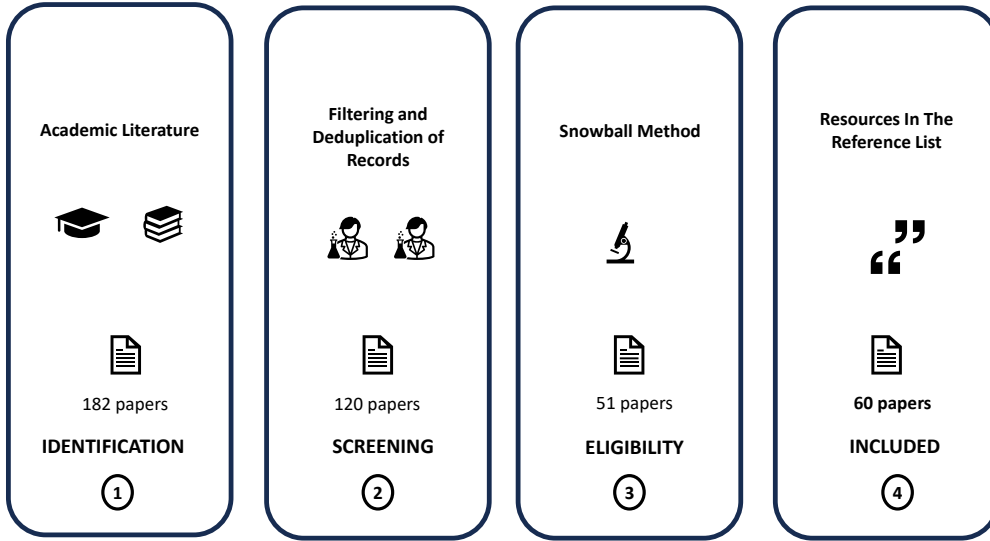

Figure 2: PRISMA flow diagram for the selection of studies on LLM-based credit risk assessment.

3.2 Research Questions

In this paper, the systematic review is structured around five research questions. Table 3 shows the research questions, motivation, and relevant section information for this paper.

4 Taxonomy of LLM-Based Credit Risk Assessment

The papers collected in this section are classified around four main categories: model architectures, data formats, interpretability mechanisms and application areas, and answers are sought to RQ1, RQ2, RQ3, and RQ4.

Table 2: Overview of 60 Selected Studies

No	Paper	Venue/Platform	Year
1	Sanz-Guerrero et al., [28]	Inteligencia Artificial (IBERAMIA)	2025
2	Dogra et al., [29]	MDPI Systems	2022
3	Dolphin et al., [30]	arXiv / Polygon.io	2024
4	Cai, [31]	IEEE ICEDCS	2024
5	Govindaraj et al., [32]	World Journal of Advanced Research Reviews	2023
6	Xie et al., [33]	arXiv (Preprint under review)	2023
7	Babaei & Giudici, [34]	Machine Learning with Applications (Elsevier)	2024
8	Loukas et al., [35]	ACM ICAIF '23 (International Conference on AI in Finance)	2023
9	Liu et al., [36]	ACM DEBAI 2024 (Digital Economy, Blockchain & AI)	2024
10	Mehedi Hasan et al., [37]	International Journal of Computer Science & Info Systems	2024
11	Charlie Luca, [38]	ResearchGate (Preprint)	2024
12	Pau Rodriguez Inserte et al., [39]	arXiv	2024
13	Linyi Yang et al., [40]	arXiv	2020
14	Jiarui Rao & Qian Zhang, [41]	International Journal of Multidisciplinary Research and Growth Evaluation	2025
15	Sungwook Yoon, [42]	International Journal of Advanced Smart Convergence	2023
16	Duanyu Feng et al., [43]	ACM (Conference acronym 'XX)	2024
17	Ayomide Joel et al., [44]	ResearchGate / Unpublished	2023
18	Jaskaran Singh Walia et al., [45]	arXiv (arXiv:2502.17011v1)	2025
19	Xue Wen Tan and Stanley Kok, [46]	ICIS (AIS Electronic Library)	2023
20	Khaoula Idbenjra et al., [47]	Elsevier – Decision Support Systems	2024
21	Zhang et al., [48]	arXiv	2023
22	Malaysha et al., [49]	arXiv	2024
23	Wu et al., [50]	CRC Working Papers	2024
24	Sideras et al., [51]	ACM (ICAIF '24)	2024
25	Fatemi et al., [52]	arXiv	2024
26	Kalluri et al., [53]	IJNRD	2024
27	Lin et al, [54]	ACM ICAIF	2024
28	Lei et al, [55]	arXiv	2025
29	Lakkaraju et al, [56]	ACM ICAIF	2023
30	Liu et al, [57]	NeurIPS Workshop	2023
31	Lopez-Lira et al., [58]	arXiv	2025
32	Xie et al., [59]	NeurIPS (Datasets & Benchmarks Track)	2023
33	Huang et al., [60]	MDPI Applied Sciences	2025
34	Moraes et al., [61]	WebMedia (Brazilian Symposium on Multimedia and the Web)	2024
35	Huang et al., [62]	arXiv	2025
36	Busireddy et al., [63]	Int. Jr. of Hum Comp. & Int.	2025
37	Babaei & Giudici, [64]	Machine Learning with Applications	2024
38	Wang et al., [65]	IEEE Transactions on Engineering Management	2025
39	Lin et al., [66]	Journal of Data, Information and Management	2025
40	Hartomo et al., [67]	IEEE Access	2025
41	Yan et al., [68]	IEEE RAAI (Robotics, Automation, and AI Conference)	2024
42	Gupta et al., [69]	JPMorgan Chase (Internal ML Research)	2024
43	Suresh et al., [70]	Educational Administration: Theory and Practice (Kuey)	2024
44	Ni et al., [71]	IEEE DOCS (Data-driven Optimization of Complex Systems)	2024
45	Bond et al., [72]	Working Paper / Preprint (University of Queensland)	2024
46	Sun et al., [73]	IEEE Access	2024
47	Sabuncuoglu et al., [74]	IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)	2025
48	Xie et al., [75]	NeurIPS (Track on Datasets and Benchmarks)	2024
49	Liu et al., [76]	IEEE 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)	2024
50	Liu et al., [77]	Finance Research Letters (Elsevier)	2025
51	Chanda and Prabhu, [78]	IEEE ICICCS Conference Proceedings	2023
52	Papasotiriou et al., [79]	ACM ICAIF '24 (AI in Finance Conference)	2024
53	Rizinski et al., [80]	IEEE Access	2024
54	Chafekar et al., [81]	Unspecified (likely arXiv or workshop preprint)	2024
55	Li et al., [82]	SSRN	2024
56	Kim et al., [83]	ACM ICAIF	2023
57	Sanz-Guerrero et al., [84]	SSRN	2024
58	Guo et al., [85]	arXiv	2023
59	Fallahgoul, [86]	SSRN (Preprint)	2025
60	Zhang et al., [87]	ACM ICAIF	2023

4.1 Model Architectures (RQ1)

Recent research on LLM-based credit risk has spanned a variety of domains, from transformer-based architectures to hybrid pipelines and domain-specific fine-tuning to efficient learning. This subsection surveys model architectures for credit risk prediction found in the literature to answer RQ1. Figure 3 shows a taxonomy of these models.

- **Encoder-Only Architectures**: Encoder-only models (RoBERTa, DistilBERT, Fin-BERT) are widely used in textual analysis-based financial classification [28]. The BERT model derives credit risk indicators and is used to integrate them with tree-based learners [28]. FinBERT is reported to perform well in financial sentiment prediction [63] and fine-tuning in encoder-based benchmarking [85]. Luca studies the BERT model for credit union use in member transactions [38]. It is also reported that BERT performs robust sentiment extraction in multilingual complex environments [83].

- **Decoder-Only Architectures**: Decoding models (GPT-3.5, GPT-4 and ChatGPT) are

Table 3: The Research Questions of the Survey

Research Question	Motivation	Section
Which LLMs are applied for credit risk assessment and how are they categorized?	The purpose is to understand LLM architectures used for credit risk applications and identify performance bottlenecks and suitability for domain-specific tasks.	4.1
What types of data sources are used in LLM-based credit risk applications?	The purpose is to evaluate how comprehensive mapping of data sources provides model generalization and fairness.	4.2
How are explainability and interpretability addressed for credit modeling in LLM-based credit risk applications?	The purpose is to evaluate how model transparency and accountability are addressed in existing studies.	4.3
In which credit risk scenarios are LLMs most commonly applied and what are the assessment metrics?	The purpose is to identify common use cases of existing studies and thus provide insight into practical relevance, dataset diversity, and reproducibility.	4.4
What are the challenges and research gaps in applying LLMs for credit risk assessment?	The purpose is to guide future research by highlighting challenges and open questions.	5

widely used in generative tasks. Loukas et al. [35] showed the performance of these models (GPT-3.5 and GPT-4) in financial intent classification with a small number of examples (8-20). Similarly, Babaei and Giudici [34, 64] report the success of GPT in low-data credit scoring scenarios.

ChatGPT has been tested by applying it to multidimensional prediction models (financial ratios, social media sentiment) [66]. Its psychological feature extraction capability is investigated and demonstrated in the GPT-LGBM framework [82]. However, Gupta et al. [69] mention limitations such as following instructions in complex documents for long-context environments in their study using GPT-4-Turbo.

- **Hybrid and Augmented Pipelines**: Some frameworks are seen to combine transform-based language modeling with traditional ML models/retrieval mechanisms. In [28], BERT and GPT are seen to be used together with XGBoost. Another work introduces a structured request pipeline in financial disclosures [30]. RAG architectures have gained attention for low-context financial data tasks (news sentiment classification) [35, 87]. All these approaches aim to improve the response base by integrating external information sources.

- **Domain-Specific Financial LLMs (FinLLMs)**: Models such as FinGPT, FinMA, ZiGong and FinLLaMA have been built to address domain variations in financial texts. FinGPT is a low-cost framework based on LoRA and Stock Price Reinforcement Learning [57]. Another study initiates the Open FinLLM Leaderboard when comparing financial LLMs and contributes to this field [54]. FinMA framework outperforms GPT-4 on credit-related tasks [59]. ZiGong [55] aims to reduce hallucinations with TracSeq using instruction tuning and temporal pruning methods. FinLMEval emphasizes that the fine-tuning ability of encoder models is superior to the zero-shot power of decoder models in data-scarce environments [85]. The Open-FinLLMs package supports textual, tabular and visual financial inputs for representing advances in multimodal credit modelling [62].

- **Parameter-Efficient LLMs and Training Techniques**: To reduce the computational cost of large models, recent studies have emphasized parameter-efficient methods. Kalluri [53] proposed a study that improves the interaction accuracy and compatibility rates with a scalable solution. In another study, Ni et al. [71] proposes QLoRA-based tuning with
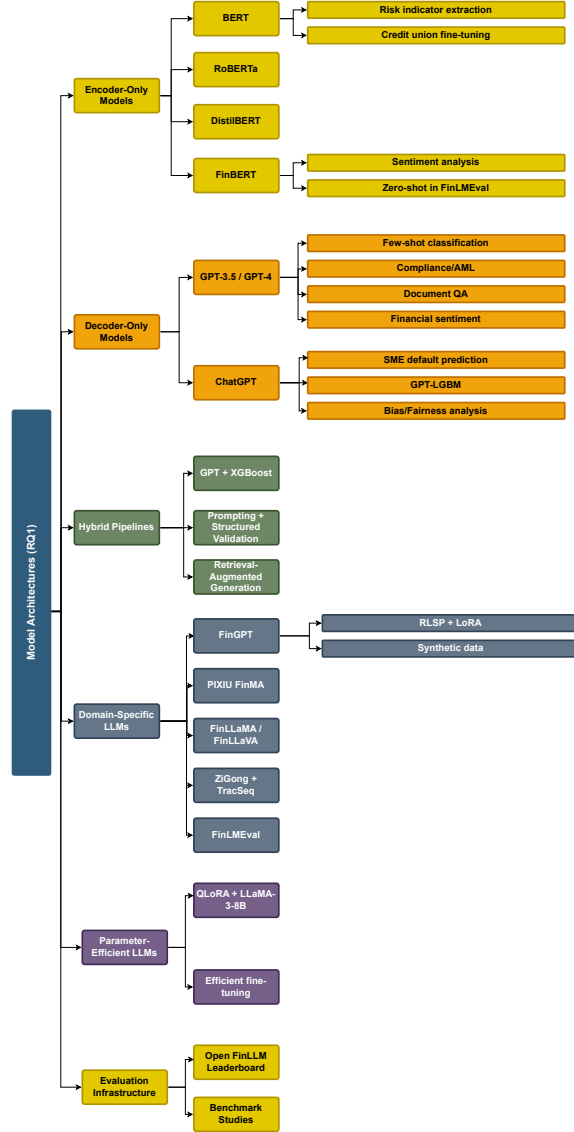
Figure 3: Taxonomy of Transformer-Based Model Architectures for LLM-Driven Credit Risk Assessment.

LLaMA-3-8B-Instruct-4bit compact models. It is emphasized that the results are better than GPT-4 for gain data.

- **Benchmarking and Evaluation Tools**: Many recent studies also aim to standardize LLM evaluation in finance. Examples include FinLLM Leaderboard [54], which allows for repeatable testing of financial benchmarks, FinLMEval [85], which provides analysis of encoder and decoder performance for fine-tuning and zero-shotting. In another study, Lakkaraju et al. [56] emphasize the concept of fairness for bias models for ChatGPT and Bard outputs.

4.2   Data Modalities (RQ2)

Recent developments have also witnessed major improvements in the data processed for LLM-focused credit risk assessment. While traditional credit modeling was done with structured tabular data, various data modalities have now emerged, such as unstructured financial text,

time series behaviors, multimodal data pipelines, and integration of synthetic data. Figure 4 shows the taxonomy of data modalities.

- **Structured Data**: Structured data such as income level and defaults are frequently used in credit scoring studies. GPT-LGBM aims to improve classification performance using this data and ChatGPT [82]. Similarly, PIXIU's FLARE benchmark and Open-FinLLMs pre-training body use structured data such as technical indicators and historical prices [59, 62].

- **Unstructured Financial Text**: Recent studies have reported that unstructured data such as credit disclosures and regulatory documents are important in capturing hidden credit indicators. Sanz-Guerrero et al. [28], Dogra et al. [29] built enhanced models with free-text disclosures and news sentiment. In other studies, Wu et al. [50] and Sideras et al. [51] examined the impact of manager and auditor comments on default accuracy.

  While ChatGPT can be used to calculate market sentiment scores using long-form news [72], it has been shown to be used effectively with GPT/BERT in the compliance profiling of AML-related documents [68]. Relatively low-resource models (such as GPT-3.5) have been shown to outperform traditional models in extracting sentiment from analyst reports [83]

- **Time-Series and Behavioral Data**: Another data type that is starting to be integrated into LLM pipelines is time series. Lei et al. [55] filter out low quality sequences in the ZiGong model with a pruning strategy and show the importance of user activity data (time series). Another case study using this data structure is FinGPT, where real-time time series data is used [57].

- **Multimodal and Hybrid Inputs**: Textual and behavioral features can be combined for more comprehensive credit modeling. Huang et al. [60] achieved high prediction accuracy using claim-based LLMs and records of 38 real-world SMEs. Another study integrates psychological features extracted from texts with financial variables in GPT-LGBM [82]. Another hybrid input example is the study by Zhang et al. [87] where LLMs are used to enrich external information by combining context-deficient financial text.

- **Synthetic and Augmented Data**: Factors affecting model performance such as data incompleteness and class imbalances can be addressed using synthetic data-generated LLM models. FinGPT is used to simulate credit records [31], another example is domain-adaptive LLM models trained on SEC filings and Reuters headlines [39]. Similarly, Feng et al. [43] proposed benchmark models for fraud and bankruptcy-related datasets, and the results show success in cross-modality learning.

4.3 Interpretability Mechanisms (RQ3)

Interpretability is an important concept in LLM-based credit risk assessment due to its fairness and reliability. Recent studies aim to increase transparency with methods such as post-hoc methods, internal model designs, instruction setting and model checking strategies. Figure 5 shows the taxonomy of interpretability mechanisms.

- **Post-Hoc Explainability**: The most commonly used post-hoc methods in this method are SHAP and LIME. Govindaraj et al. [32] proposed a study to visualize local global feature contributions by applying SHAP and attention heatmaps. Liu et al. [36] conducted a study combining SHAP and LIME outputs with ChatGPT to make them human interpretable.In [50], LIME's efficiency in credit assessments is examined, while in [67] SHAP is used with TabTransformer for class skew compensation and interpretability.
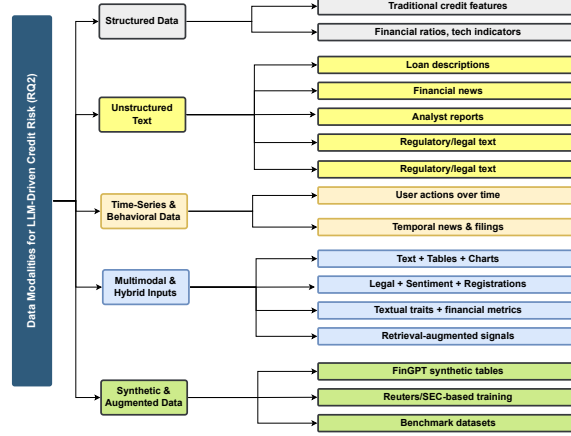
Figure 4: Taxonomy of Data Modalities Utilized in LLM-Based Credit Risk Systems.

- **Chain-of-Thought and Prompt- Level Explanations**: Recent work aims to enable models to generate explanations directly in the inference pipeline. Dolphin et al. [30] use chain-of-thought prompts to provide interpretability in sentiment classification scenarios, while in [35] they use backoff to improve prediction fixation and tractability. Fatemi et al. [52] show that interpretability can be increased by improving the model's internal reasoning through instruction tuning.

- **Intrinsically Interpretable Model Designs**: Some studies propose transparent architectures to reduce the dependency on external explainability tools. One of these models is Logit Leaf, which integrates LLM outputs and segmentation trees, as proposed by Idben-jra et al. [47]. Another study introduces FinBERT-XRC, which can interpret risks at the word and sentence level [46] . Li et al. [82] reported that the personality traits extracted by GPT-LGBM are directly interpretable (without post-hoc XAI).

- **Robustness and Hallucination Mitigation**: One of the factors that directly affects interpretability is model robustness. To increase model robustness and address hallucinations, in [55], the authors propose a model called TracSeq. In another study, the authors aim to improve interpretability by combining gain deltas and the QLoRA model.

- **Fairness, Auditing, and Reproducibility**: Other concepts that are directly related to interpretability are fairness and reproducibility. In [56] ISIP and ISA metrics are used to check model fairness. Gupta et al. [69] emphasize the importance of F1 score and confidence intervals (CI) metrics for explainability checks. Additionally, Lin et al. [54] advocate benchmarking with the FinLLM Leaderboard to standardize interpretability assessment.

- **Ethical, Regulatory, and Theoretical Dimensions**: For reliable interpretability, ethics should also be taken into account. Yan et al. [68] highlights six basic dimensions of compliance to evaluate LLMs in critical applications: correctness, fairness, privacy, robustness, security, and ethics. Another paper [86] discusses how to extend the Markowitz framework to provide attention-based interpretability and optimization.

4.4 Application Domains (RQ4)

LLMs have recently been applied in a wide range of areas, including traditional credit scoring, fraud detection, sentiment prediction, robo-advisory, etc. Figure 6 shows the Taxonomy of Application Domains for LLMs in Credit Risk Assessment.
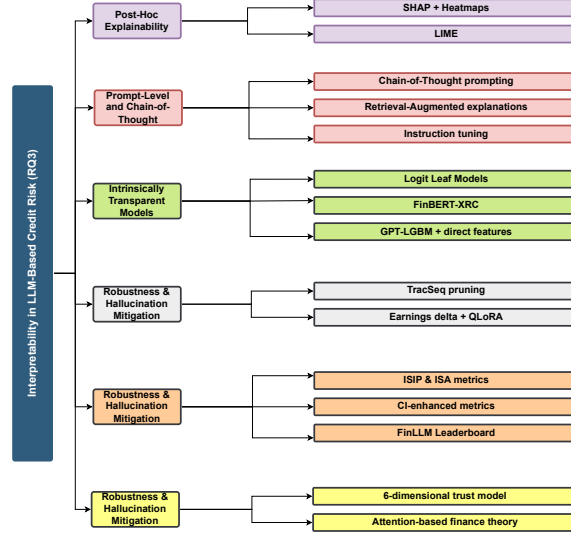
Figure 5: Taxonomy of Interpretability Mechanisms for LLM-Based Credit Models.

- **Retail and SME Credit Scoring**: When the literature is examined, it is seen that LLMs are used in various financial application areas such as peer-to-peer (P2P) lending and SME financing decisions. In their studies, Sanz-Guerrero et al. [28] and Babaei & Giudici [34] use credit disclosures for borrower creditworthiness classification. In [42], SME technology loans are evaluated with a GPT-based credit assessment mechanism. In another study, [60], claims-based assessments are applied to non-financial data for SME default prediction. In [67], the SHAP-TabTransformer model is applied to SME credit classification. Li et al. [82] combine structured financial data with extracted personality traits for credit assessment in their proposed GPT-LGBM framework.

- **Financial News, Sentiment, and Market Signals**: LLMs performed well in extracting market sentiment and event triggers even on unstructured financial data. Dolphin et al. [30] and Dogra et al. [29] used LLMs to predict risk from headlines with event triggers. In [72] daily S&P 500 sentiment scores were generated using GPT-3.5, while in another study [83] analyst tone was measured. Seshakagari et al [63] used GPT-4 to classify financial sentiment in a changing macroeconomic environment.

- **Customer Service and Banking Operations**: Some studies in the literature focus on intent recognition and service personalization with LLMs. Loukas et al. [35] performed service request classification with the Banking77 dataset, while in [70] customer sentiment analysis was used to support banking service personalization and credit decisions.

- **Fraud Detection and Anti-Money Laundering (AML)**: FinGPT and ZiGong are two models used in fraud detection in credit and banking transactions [31, 55]. In [44] credit unions predict fraud and default, while Yan et al. [68] examine the performance of GPT/BERT-based models on AML, sanctions screening, and suspicious activity.

- **Investment, Trading, and Asset Management**: Walia et al. [45] apply LLMs to bond yield prediction and extend the use of LLMs in finance. In [54] LLMs are used for SEC filing interpretation and sentiment-based trading, while Liu et al. [57] evaluate the performance of FinGPT in robo-advisory and algorithmic trading bots. Other studies test the capabilities of LLMs using Open-FinLLMs in multimodal investment reasoning [62]. Ni et al. [71] report that QLoRA outperforms GPT-4 in stock prediction.

- **Taxonomy Building and Transaction Analysis**: In [61], LLMs are used to create financial taxonomy and explain transaction data to perform credit analyses, thus strengthening banking behaviors in credit and KYC (Know Your Customer) scenarios.

- **Supply Chain and Sector-Specific Credit Evaluation**: In [65], LLMs use bid data and operational descriptions to evaluate China's green transportation sector SME loans, helping non-specialist analysts understand feasible financing targets and strengthening the accessibility of LLMs.

- **Early Warning Systems and Multidimensional Risk Ratings**: ChatGPT was used for credit risk estimation scenarios where data is incomplete, such as start-ups [66].This provides early warning and adaptive credit scoring capabilities for variable domains and applications.
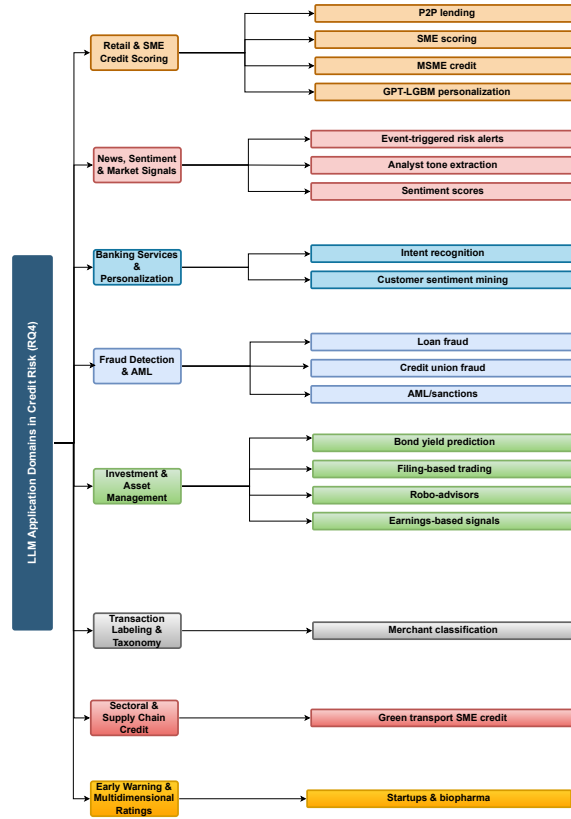

Figure 6: Taxonomy of Application Domains for LLMs in Credit Risk Assessment.

5 Research Gaps and Future Directions (RQ5)

Although significant advances have been made in LLM-based financial applications (credit risk estimation, etc.), there are still research gaps and limitations that need to be investigated. Figure 7 illustrates these limitations.

- **Interpretability Gaps**: LLMs applied in finance generally remain black boxes because they rely on post hoc methods (such as SHAP and attention maps) for explainability [32, 36]. Causal or counterfactual reasoning is almost absent in LLMs applied in finance,

and therefore it is difficult to understand the real decision mechanism of the model. Although recent studies ([40]) propose counterfactual reasoning, its adoption remains minimal. Furthermore, although research is ongoing on LLMs being able to perform automatic classification ([61]), they are still not widely used in taxonomy studies.

- **Reproducibility and Robustness Limitations**: Few literature studies have performed robustness testing. Most experiments appear to use small or parsimonious datasets [33, 35]. Xie et al. [59] reported that FinMA struggles with reasoning tasks and that models trained with human-like instructions are vulnerable.

- **Bias, Fairness, and Hallucination Risks**: The literature reports that LLMs are prone to bias against demographic characteristics such as race, gender, and age [36, 56]. Furthermore, LLMs can produce false information, hallucinations, that are very close to the truth, and these remain a major threat to credit risk estimation [37]. Few frameworks in the literature target fairness and hallucination reduction [68].

- **Efficiency and Model Scaling**: Very few of the studies consider metrics such as latency, inference cost, and hardware performance [35, 39]. Although models like FinLLaMA are calculated considering their small footprint, there is still a research gap in this area [62].

- **Evaluation Benchmark Deficiencies**: Due to the diversity of datasets and scenarios, it is difficult to directly compare the performance of LLMs, a shortcoming noted in some literature studies (such as the assessment of the trade-off between interpretability and complexity) [54, 86].

- **Integration of Behavioral and External Signals**: Signals from unstructured data (news, social media content, etc.) can be incorporated into credit scoring models and used for forecasting purposes. Although theoretically potential, behavioral and external signals have not yet been widely integrated.
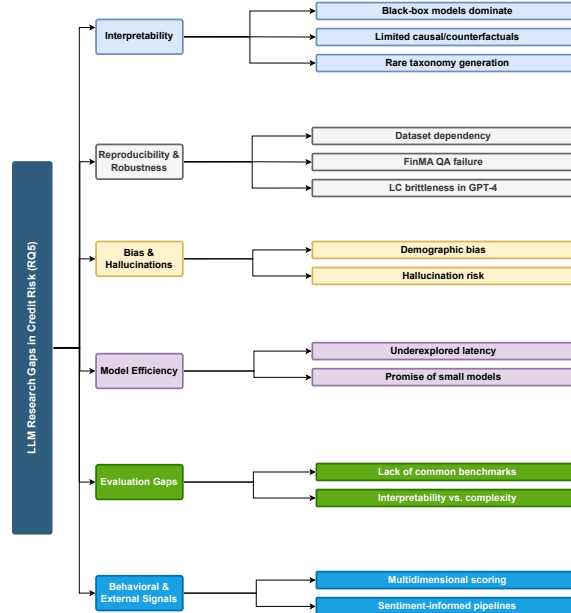

Figure 7: Taxonomy of Research Gaps and Future Directions in LLM-Driven Credit Risk Research.

Future Directions: Based on the research gaps described above, the following areas stand out for future researchers:

- **Low-Cost Models**: Compact models such as QLoRA-fine-tuned LLaMA ([71]) variants can be investigated to develop cost-aware credit scoring models.

- **Evaluation Criteria**: Fairness-focused task-specific evaluation criteria such as context length decay and hallucination robustness can be developed.

- **Fairness and Trust Protocols**: In future research, auditing tools and trust frameworks can be created for consumer finance.

- **Sentiment-Driven Credit Signals**: New studies can be conducted by including concepts such as investor behavior and textual sentiment for loan pricing.

- **Behavioral Personalization**: New LLM-based studies can develop digital banking applications with intent-aware and emotion-aware personalization.

- **Regulatory Consistency and Compliance**: Studies can be conducted on making LLM models compliant with ethical and legal standards in AML applications and anti-bias applications.

6 Conclusions

This paper presents the first systematic review and taxonomy of Large Language Model (LLM) based credit risk assessment approaches. The most relevant 60 peer-reviewed studies published between 2020 and 2025 are analyzed using the PRISMA methodology and a structured taxonomy is presented along four main dimensions: model architectures, data formats, interpretability mechanisms, and application domains. The findings confirm that although only coding and domain-specific FinLLMs are frequently used in this field, recent trends in hybrid pipelines, parameter-efficient tuning (e.g. QLoRA), and multimodal data integration are also used in credit scoring. Although SHAP and LIMA (post-hoc) are prevalent interpretability techniques, there is increasing interest in intrinsic and demand-based explanations. Apart from all these developments, there is still a need for further research in the areas of reproducibility, fairness control, robustness to hallucinations, and standardized assessment. We hope that this systematic review and taxonomy study will be a reference paper for researchers in transparent, reliable and field-compatible LLM-based financial risk modeling.

References

[1] Zhang Nana, Wei Xiujian, and Zhang Zhongqiu. Game theory analysis on credit risk assessment in e-commerce. *Information Processing & Management*, 59(1):102763, 2022.

[2] Jan Roeder, Matthias Palmer, and Jan Muntermann. Data-driven decision-making in credit risk management: The information value of analyst reports. *Decision Support Systems*, 158:113770, 2022.

[3] Ji-Won Kang and Sun-Yong Choi. Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors. *Electronics*, 14(6):1090, 2025.

[4] Muhammed Golec, Yaser Khamayseh, Suhib Bani Melhem, and Abdulmalik Alwarafy. Llm-driven apt detection for 6g wireless networks: A systematic review and taxonomy. *arXiv preprint arXiv:2505.18846*, 2025.

[5] Olamilekan Shobayo, Sidikat Adeyemi-Longe, Olusogo Popoola, and Bayode Ogunleye. Innovative sentiment analysis and prediction of stock price using finbert, gpt-4 and logistic regression: A data-driven approach. *Big Data and Cognitive Computing*, 8(11):143, 2024.

[6] Daniel Staegemann, Christian Haertel, Christian Daase, Matthias Pohl, Mohammad Abdallah, and Klaus Turowski. A review on large language models and generative ai in banking.

[7] Luke Lee. Enhancing financial inclusion and regulatory challenges: A critical analysis of digital banks and alternative lenders through digital platforms, machine learning, and large language models integration. *arXiv preprint arXiv:2404.11898*, 2024.

[8] Satyadhar Joshi. Gen ai for market risk and credit risk learn agentically powered gen ai; gen ai agentic framework for financial risk management. *Gen AI Agentic Framework for Financial Risk Management (January 15, 2025)*, 2025.

[9] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024.

[10] Muhunthan Jayanthakumaran, Nagesh Shukla, Biswajeet Pradhan, and Ghassan Beydoun. A systematic review of sentiment analytics in banking headlines. *Decision Analytics Journal*, page 100584, 2025.

[11] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

[12] Elias Zavitsanos, Eirini Spyropoulou, George Giannakopoulos, and Georgios Paliouras. Machine learning for identifying risk in financial statements: A survey. *ACM Computing Surveys*, 57(9):1–37, 2025.

[13] Layla Abdel-Rahman Aziz and Yuli Andriansyah. The role artificial intelligence in modern banking: an exploration of ai-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. *Reviews of Contemporary Business Analytics*, 6(1):110–132, 2023.

[14] Satyadhar Joshi. A comprehensive review of gen ai agents: Applications and frameworks in finance, investments and risk domains. 2025.

[15] Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management*, 51(2), 2024.

[16] Satyadhar Joshi. Review of gen ai models for financial risk management: Architectural frameworks and implementation strategies. *Available at SSRN 5239190*, 2025.

[17] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. Large language models in finance (finllms). *Neural Computing and Applications*, pages 1–15, 2025.

[18] Farhina Sardar Khan, Syed Shahid Mazhar, Kashif Mazhar, Dhoha A. AlSaleh, and Amir Mazhar. Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, 58(8):232, 2025.

[19] Amin Karami and Chukwuemeka Igbokwe. The impact of big data characteristics on credit risk assessment. *International Journal of Data Science and Analytics*, pages 1–21, 2025.

[20] Noguer I Alonso et al. Large language models in finance: Reasoning. *Large Language Models in Finance: Reasoning (December 08, 2024)*, 2024.

[21] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023.

[22] Kausik Lakkaraju, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath Muppasani, and Biplav Srivastava. Can llms be good financial advisors?: An initial study in personal decision making for optimized outcomes. *arXiv preprint arXiv:2307.07422*, 2023.

[23] Steve Samson and Jitendra Rout. Comparative analysis of large language models adaptability to detect sentiments in financial domain.

[24] Adetoyese Omoseebi, Akorede Jhon, and Winner Olabiyi. Large language models (llms) for financial security. 2024.

[25] Julian Junyan Wang and Victor Xiaoqi Wang. Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *arXiv e-prints*, pages arXiv–2503, 2025.

[26] David Krause. Large language models and generative ai in finance: an analysis of chatgpt, bard, and bing ai. *Bard, and Bing AI (July 15, 2023)*, 2023.

[27] Jun Xu. Genai and llm for financial institutions: A corporate strategic survey. *Available at SSRN 4988118*, 2024.

[28] Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *arXiv preprint arXiv:2401.16458*, 2024.

[29] Varun Dogra, Fahd S Alharithi, Roberto Marcelo Álvarez, Aman Singh, and Abdulrahman M Qahtani. Nlp-based application for analyzing private and public banks stocks reaction to news events in the indian stock exchange. *Systems*, 10(6):233, 2022.

[30] Rian Dolphin, Joe Dursun, Jonathan Chow, Jarrett Blankenship, Katie Adams, and Quinton Pike. Extracting structured insights from financial news: An augmented llm driven approach. *arXiv preprint arXiv:2407.15788*, 2024.

[31] Yuanxi Cai. Overcoming data limitations in credit risk assessment with fingpt-generated synthetic data. In *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 137–141. IEEE, 2024.

[32] Vasanthi Govindaraj, Humashankar Vellathur Jaganathan, and P Prakash. Explainable transformers in financial forecasting. 2023.

[33] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

[34] Golnoosh Babaei and Paolo Giudici. Gpt classifications, with application to credit lending. *Machine Learning with Applications*, 16:100534, 2024.

[35] Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 392–400, 2023.

[36] Zhengping Liu, Hailing He, Lieping Zhang, and Chen Peng. Leveraging xai in prompt-based chatgpt for financial decision support. In *Proceedings of the International Conference on Digital Economy, Blockchain and Artificial Intelligence*, pages 255–259, 2024.

[37] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255, 2023.

[38] Charlie Luca. Optimizing large language models for financial risk assessment in credit unions. 2024.

[39] Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. Large language model adaptation for financial sentiment analysis. *arXiv preprint arXiv:2401.14777*, 2024.

[40] Linyi Yang, Eoin M Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. Generating plausible counterfactual explanations for deep transformers in financial text classification. *arXiv preprint arXiv:2010.12512*, 2020.

[41] Jiarui Rao and Qian Zhang. Deep learning with llm: A new paradigm for financial market prediction and analysis. 2025.

[42] Sungwook Yoon. Design and implementation of an llm system to improve response time for smes technology credit evaluation. *International journal of advanced smart convergence*, 12(3):51–60, 2023.

[43] Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*, 2023.

[44] AYOMIDE JOEL. Optimizing large language models for financial risk assessment in credit unions.

[45] Jaskaran Singh Walia, Aarush Sinha, Srinitish Srinivasan, and Srihari Unnikrishnan. Predicting liquidity-aware bond yields using causal gans and deep reinforcement learning with llm evaluation. *arXiv preprint arXiv:2502.17011*, 2025.

[46] Xue Wen Tan and Stanley Kok. Explainable risk classification in financial reports. *arXiv preprint arXiv:2405.01881*, 2024.

[47] Khaoula Idbenjra, Kristof Coussement, and Arno De Caigny. Investigating the beneficial impact of segmentation-based modelling for credit scoring. *Decision Support Systems*, 179:114170, 2024.

[48] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*, 2023.

[49] Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Almujaiwel, Ismail Berrada, and Houda Bouamor. Arafinnlp 2024: The first arabic financial nlp shared task. *arXiv preprint arXiv:2407.09818*, 2024.

[50] Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. Unleashing the power of text for credit default prediction: Comparing human-generated and ai-generated texts. *SSRN Electronic Journal. https://doi. org/10.2139/ssrn*, 4601317, 2023.

[51] Andreas Sideras, Konstantinos Bougiatiotis, Elias Zavitsanos, Georgios Paliouras, and George Vouros. Bankruptcy prediction: Data augmentation, llms and the need for auditor's opinion. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 453–460, 2024.

[52] Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. A comparative analysis of instruction fine-tuning llms for financial text classification. *arXiv preprint arXiv:2411.02476*, 2024.

[53] Kartheek Kalluri. Scalable fine-tunning strategies for llms in finance domain-specific application for credit union, 2024.

[54] Shengyuan Colin Lin, Felix Tian, Keyi Wang, Xingjian Zhao, Jimin Huang, Qianqian Xie, Luca Borella, Matt White, Christina Dan Wang, Kairong Xiao, et al. Open finllm leaderboard: Towards financial ai readiness. *arXiv preprint arXiv:2501.10963*, 2025.

[55] Yu Lei, Zixuan Wang, Chu Liu, and Tongyao Wang. Zigong 1.0: A large language model for financial credit. *arXiv preprint arXiv:2502.16159*, 2025.

[56] Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. Llms for financial advisement: A fairness and efficacy study in personal decision making. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 100–107, 2023.

[57] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Data-centric fingpt: Democratizing internet-scale data for financial large language models. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023.

[58] Alejandro Lopez-Lira, Jihoon Kwon, Sangwoon Yoon, Jy-yong Sohn, and Chanyeol Choi. Bridging language models and financial analysis. *arXiv preprint arXiv:2503.22693*, 2025.

[59] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

[60] Haonan Huang, Jing Li, Chundan Zheng, Sikang Chen, Xuanyin Wang, and Xingyan Chen. Advanced default risk prediction in small and medum-sized enterprises using large language models. *Applied Sciences*, 15(5):2733, 2025.

[61] Daniel de S Moraes, Polyana B da Costa, Pedro TC Santos, Ivan de JP Pinto, Sérgio Colcher, Antonio JG Busson, Matheus AS Pinto, Rafael H Rocha, Rennan Gaio, Gabriela Tourinho, et al. Tagging enriched bank transactions using llm-generated topic taxonomies. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 267–274. SBC, 2024.

[62] Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024.

[63] Haranadha Reddy Busireddy Seshakagari, Aravindan Umashankar, T Harikala, L Jayasree, and Jeffrey Severance. Dynamic financial sentiment analysis and market forecasting through large language models. *International Journal of Human Computations & Intelligence*, 4(1):397–410, 2025.

[64] Golnoosh Babaei and Paolo Giudici. Gpt classifications, with application to credit lending. *Machine Learning with Applications*, 16:100534, 2024.

[65] Jiaxing Wang, Guoquan Liu, Yang Cheng, Xiaobo Xu, and Zhongyun Li. Leveraging internet-sourced text data for financial analytics in supply chain finance: A large language model-enhanced text mining workflow. *IEEE Transactions on Engineering Management*, 2025.

[66] Jinlin Lin, Sirui Lai, Hao Yu, Rui Liang, and Jerome Yen. Chatgpt based credit rating and default forecasting. *Journal of Data, Information and Management*, pages 1–24, 2025.

[67] Kristoko Dwi Hartomo, Christian Arthur, and Yessica Nataliani. A novel weighted loss tabtransformer integrating explainable ai for imbalanced credit risk datasets. *IEEE Access*, 2025.

[68] Yuqi Yan, Tiechuan Hu, and Wenbo Zhu. Leveraging large language models for enhancing financial compliance: A focus on anti-money laundering applications. In *2024 4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, pages 260–273. IEEE, 2024.

[69] Lavanya Gupta, Saket Sharma, and Yiyun Zhao. Systematic evaluation of long-context llms on financial concepts. *arXiv preprint arXiv:2412.15386*, 2024.

[70] Dr M Suresh, G Vincent, C Vijai, M Rajendhiran, M Com, AH Vidhyalakshmi, and S Natarajan. Analyse customer behaviour and sentiment using natural language processing (nlp) techniques to improve customer service and personalize banking experiences. *Educational Administration: Theory And Practice*, 30(5):8802–8813, 2024.

[71] Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, Panfeng Li, Shiyao Zhang, Qifu Yin, Yuanqing Wang, and Yuxi Chan. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 909–915. IEEE, 2024.

[72] Shaun A Bond, Hayden Klok, and Min Zhu. Large language models and financial market sentiment. *Available at SSRN 4584928*, 2023.

[73] Tiejiang Sun, Jingyun Yang, Jiale Li, Jiaying Chen, Mingyue Liu, Li Fan, and Xukang Wang. Enhancing auto insurance risk evaluation with transformer and shap. *IEEE Access*, 2024.

[74] Alpay Sabuncuoglu and Carsten Maple. Identifying representation bias in large language models used in financial sentiment analysis. In *2025 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CiFer)*, pages 1–7. IEEE, 2025.

[75] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

[76] Zhiyi Liu, Kai Zhang, Yejie Zheng, and Zheng Sun. Research on the application methods of large language model interpretability in fintech scenarios. In *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 526–531. IEEE, 2024.

[77] Yingnan Liu, Ningbo Bu, Zhiqiang Li, Yongmin Zhang, and Zhenyu Zhao. At-fingpt: Financial risk prediction via an audio-text large language model. *Finance Research Letters*, 77:106967, 2025.

[78] Rajat Chanda and Sandeep Prabhu. Secured framework for banking chatbots using ai, ml and nlp. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 60–65. IEEE, 2023.

[79] Kassiani Papasotiriou, Srijan Sood, Shayleen Reynolds, and Tucker Balch. Ai in investment analysis: Llms for equity stock ratings. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 419–427, 2024.

[80] Maryan Rizinski, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov. Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access*, 12:7170–7198, 2024.

[81] Chafekar Talha, Hussain Aafiya, and Cheong Chon In. Understanding behaviour of large language models for short-term and long-term fairness scenarios. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 52–61, 2023.

[82] Li Yu, Xuefei Bai, and Zhiwei Chen. Gpt-lgbm: A chatgpt-based integrated framework for credit scoring with textual and structured data. *Available at SSRN 4671511*, 2023.

[83] Seonmi Kim, Seyoung Kim, Yejin Kim, Junpyo Park, Seongjin Kim, Moolkyeol Kim, Chang Hwan Sung, Joohwan Hong, and Yongjae Lee. Llms analyzing the analysts: Do bert and gpt extract more value from financial analyst reports? In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 383–391, 2023.

[84] Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in peer-to-peer lending. *Available at SSRN 4979155*.

[85] Yue Guo, Zian Xu, and Yi Yang. Is chatgpt a financial expert? evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*, 2023.

[86] Hasan Fallahgoul. Beyond black-box ai: A theory of interpretable transformers for asset pricing. *Available at SSRN*, 2025.

[87] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 349–356, 2023.