Mathematical Modelling of Engineering Problems
Vol. 12, No. 9, September, 2025, pp. 3237-3246
Journal homepage: http://iieta.org/journals/mmep

Automated Compliance Checking of Unstructured Data Using Large Language Models and Langchain

Yash Buddhadev[1], Harsh Chitaliya[1], Vatsal Kotha[2], Darshana Sankhe[3], Paresh Nasikkar[4*], Pratik Kanani[1], Mrunal Rane[1]

[1] Department of Artificial Intelligence and Data Science, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India
[2] Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India
[3] Department of Electronics and Telecommunication, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India
[4] Department of Electronic and Telecommunication, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), Pune 412115, India

Corresponding Author Email: paresh.nasikkar@sitpune.edu.in

https://doi.org/10.18280/mmep.120927

Received: 23 June 2025
Revised: 12 August 2025
Accepted: 18 August 2025
Available online: 30 September 2025

Keywords:
automation of financial audits, compliance checking, different types of audits, Large Language Models (LLMs) and Langchain, PDF text extraction, unstructured financial data

ABSTRACT

Manual compliance audits on scanned or unstructured documents are often time-consuming as well as error-prone. This paper proposes an artificial intelligence-based solution using Optical Character Recognition (OCR), Large Language Models (LLMs) and Langchain to leverage technology in order to automate the entire procedure of compliance checking. It follows two primary objectives: (1) an annual compliance verification procedure for various types of documents: such as images, scanned copies and PDF files of annual/statutory audits, that in return produces a highlighted downloadable report indicating the status of compliance and highlighting the irregularities, and (2) an enhanced procedure for internal and concurrent audits which appends all compliance data into a singular CSV file, enabling real-time audit trails at a one-stop location. Advanced OCR and Natural Language Processing (NLP) tools like Facebook's Bidirectional and Auto-Regressive Transformer (BART) and Microsoft's Transformer-based Optical Character Recognition (TrOCR) are thus utilized to convert unstructured data so that non-compliant areas are easily identified and flagged. The system reduces the audit time of a mid-sized bank by almost 45% to 50% with an accuracy rate of upto 95%, which means that it would reduce the cost by up to 35% over three years compared to existing traditional methods.

1. INTRODUCTION

While exploring the idea of pursuing an AI-based approach for possibly automating the lengthy and manual auditing and compliance procedure, we came across a couple of studies around the same.

1.1 Motivation and background

Finance fundamentally, is a big umbrella that absorbs a broad spectrum having various sub-domains, crucially contributing to the world's economics, also market situation since decades [1]. Right from personal finance to markets and equity, to regulatory technology (RegTech), these sectors make up a solid portion of the financial world as we see today. For startups and existing businesses alike, keeping any of these financial arenas aside can have serious and far-reaching implications on their existence as well as their market positions respectively [2].

From these deep dimensions of finance as seen in Figure 1, one such revolutionizing force is FinTech, where financial institutions adapt technology with its advancements. This field indicates deep shift from the traditional means of financial services [3]. In 2023, the market size of FinTech is enormous with different subdomains carving out niches for themselves and contributing towards the overall market valuation numbers [4].

Figure 1. Major domains in finance

From these broad alleyways of finance as seen in Figure 1, we focus on the RegTech sector. This field indicates deep

change from the traditional means of financial services through technology [3]. In 2023, the market size of the enormous FinTech domain is humongous with different subdomains carving out niches for themselves and contributing towards the overall market valuation, having RegTech and AuditTech both as major contributors [4].

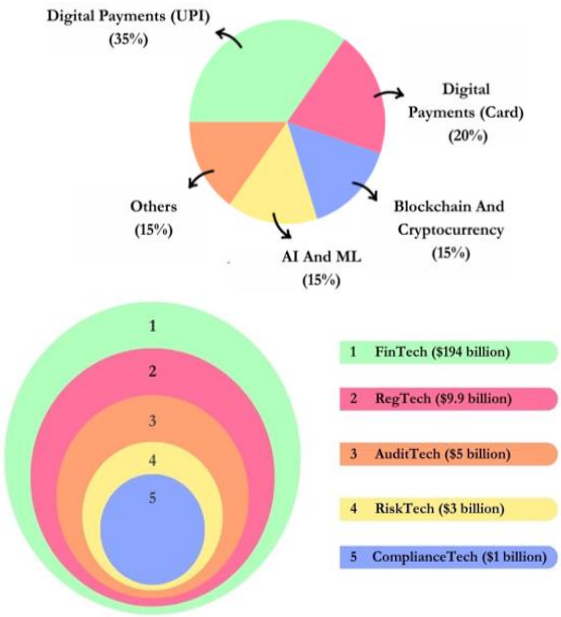

Figure 2. Market size of sub-domains in finance

In this paper, we focus on the very critical sector of the FinTech landscape: RegTech or simply put, RegTech. As can be seen from the market size diagram Figure 2, RegTech occupies a vast part of the FinTech market with an estimated value of $9.9 billion in 2023 [5]. Under the domain of RegTech, major sections include auditing and compliance, which are significant for the integrity as well as legality of any financial operation.

Understanding the scenario so far, we aimed to identify a problem and a way to go about solving it, with our innovative approach:

1.2 Identified problem and potential solution

There are various types of audits in auditing that have been carried out to maintain financial integrity and comply with regulations [6]. Audits can be mainly categorized into three types, which include internal audits, concurrent audits, and statutory audits as indicated in Figure 3. Internal audits involve a company's own staff or an independent internal team evaluating internal controls, risk management, and governance processes. Concurrent audits are live examinations of bank or financial house financial transactions to detect and correct faults in time. Statutory audits are legal compulsions of having an external auditor examine a company's financial statements to ensure their accuracy and compliance with the legal and regulatory standards [7].

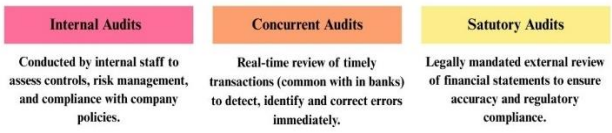

Figure 3. Types of audits

Traditionally, the audit process entailed much manual labor when dealing with unstructured information, like PDFs, scanned images, or handwritten documents [8]. This challenge is specifically covering the context of internal and concurrent audits that are recurring over time, this also involves processing giant chunks of data against different set of guidelines and compliance checkpoints [9]. Advanced unstructured data processing techniques will be used to demonstrate the automation and simplification of compliance checking across various types of audits in our research on challenges faced by banks in auditing processes [10].

The proposed approach holds the potential of bringing revolution into the auditing process for this sector and achieve high efficiency improvements, accuracy, and cost-effectiveness within the banking sector catering to all 3 types of audits mentioned [11]. This works wonders as it covers all the dimensions and shortcomings such as long, stressful hours of manual work required to be done with high concentration, attention to detail for the minute edge cases, overall uniformity in the compliance-check procedure. The automated approach we describe below can significantly reduce the time spent on manually reviewing documents; projected to be between 20% and 75% off [10]. That means an average of 20 hours a week saved for the average team of auditors [11]. In terms of productivity, the system could read and process up to 1,000 pages per hour of documentation, a far cry from the average manual processing rate of 50-100 pages per hour [12]. Even from a cost perspective, the system is expected to result in a reduction in audit cost by 40% for a mid-sized bank over a period of three years [11]. Our projections indicate that in a large bank institution, full implementation would lead to an increase of 30% in annual audits completed, and a 50% reduction in the time taken for each audit [10] and estimated annual savings to the tune of $2 million for a bank with over $50 billion in assets [9].

2. LITERATURE REVIEW

2.1 Optical Character Recognition (OCR) for document processing

Unstructured data, particularly from IoT devices, is constantly generated but frequently not well managed for long-term reuse. Machine learning techniques provide the foundation for turning such unstructured data into organized representations in order to improve information extraction [13]. The proper processing of such data and the extraction of insights from it pose significant challenges.

The use of audit data analytics in auditing practice today is limited to its use by large international accounting firms in the audit exam process of financial statements. Auditing data analytics consists of analysis and graphical presentation of audit related data to identify trends, highlight anomalies, and develop meaningful insights [14]. Detailed information on the Auditing Data Analytics (ADA) practices of large firms was collected using 63-minute interviews.

Text mining is essential to auditing because it provides valuable insight into emerging behaviors from unstructured data in the search for patterns of emerging behaviors, used for analysis despite the limitations of any customary structured data research study [15]. Unstructured data management solutions include ones to manage content management systems, Hadoop, Massively Parallel Processing (MPP) data

warehouses, NoSQL database management systems such as MongoDB, in-memory database systems such as Systems, Applications, and Products (SAP) Hana, and even log monitor tools which can parse massive lots of data from unstructured sentential logs [16]. Big data analytic software such as HDFS, MapReduce, and HBase can transform unstructured data into usable structured data types with processing capabilities, including parallel processing, allowing auditing to go where it has not gone before relative to customary analytics and at the same time facilitating discipline change because of the scope of data [17].

Research notes on OCR are plentiful, and much of the OCR work has focused on automatic extraction on the already established parts of documents. There are systems compared that either are based on PDF in Figure 2, one compared deep figures, and the other two methods being segmentation-based or detection-based compared across recall, precision, and F1 varies overall. The research that used PDF in Figure 2 to analyze OCR showed the best overall performance [18]. Many open-source projects are also taking the different approaches to make parsing free-form PDFs easier. The text areas are outlined with bounding boxes, with transparency lowered in the PyMuPDF library so the user can then tidy extracted text before moving on to effectively analysis [19].

Critical gap analysis: Although the approaches are strong in terms of digitisation and structuring data, they seldom include using compliance rules during extraction, or only a few moments after, meaning effective auditing cannot take place in a timely fashion.

2.2 Large Language Models (LLMs) for compliance interpretation

Recent developments and advancements with LLMs have displayed powerful abilities to quickly acquire, and comprehend, very complex, unstructured financial information. For example, GPT-4 (96.8%), Claude 2 (93.7%) and Gemini (69%) [20]. The accuracy should improve due to the improved accuracy with the Retrieval-Augmented Generation (RAG) where the model output will have to be substantively tied and referenced back to an external verified data source [21].

Domain specific adaptations of models, such as FinGPT, have adapted multiple models for finance and achieved accuracy scores ranging as high as 0.887 [22]. FETILDA's framework has adapted multimodal and chunking methods including FinBERT and Longformer to handle token length issues caused by lengthy financial texts by a significant margin [23]. GPT-4 has even been extended to large datasets of financial text with some human-in-the-loop reliability checks [24].

Besides financial reporting, LLMs using graph-based techniques have also assisted by extracting and acquiring Environmental, Social and Governance (ESG) information from sustainability reports, and had higher efficiency than the earlier Open Information Extraction method [25].

Critical gap analysis: The literature tells us that LLMs demonstrate a significant amount of semantic knowledge within the compliance setting. The literature has, however, observed that LLMs are hardly used as part of compliance checks on handwritten or scanned documents in OCR pipelines.

2.3 Previous RegTech systems

RegTech employs mechanisms of automation to deliver efficiencies in compliance. The Basel Accord continues to change but mechanisms were developed called costly state verification, deterministic auditing, stochastic auditing, and dynamic capital schedules is all functioning to help financial entities, by modifying capital requirements through compliance performance [26].

ADA and big data techniques have also enabled anomaly detection and fraud risk detection, and improved audit reliability [14, 15]. The exploration of automation has also been documented relating to bank auditing. Robotic process automation and artificial intelligence models have been identified, and put into practice, in the area of bank auditing with measurable efficiencies [7-9].

3. RESEARCH GAP

- Current scenario clearly depicts a shortage of a one-place approach that can be used to build a system catering to all types of audits for private as well as governmental institutions.
- There is no standard framework for classifying compliance checks as per the guidelines in an automated manner.
- Current market scenario depicts the need for scalable and integrated solutions, that have a tech-first approach to leverage and utilize the advancements in Natural Language tasks with the updated LLMs in the areas of managing volumes and complexities in huge audit trails, to work seamlessly while also maintaining a standard procedure, having decent accuracy metrics to detail while checking compliance on the required set of regulations.
- Banks usually invest a lot of time in the manual procedure of checking audit trails, that can go up to a few months in most cases, having large teams working for a singular trail status, and still are more exposed to manual errors like calculations or missing one of the certain edge cases, or even lacking a uniform approach of inspection while handling audit reports that include thousands of pages to be processed for each line.

4. PROPOSED SOLUTION

Legacy audit procedures face major challenges associated with manual document identification and verification, and tracking compliance checks. These challenges create bottlenecks, increase the likelihood of errors, and limit scalability. To overcome these challenges, this research proposes an automated auditing system using advanced OCR, Natural Language Processing (NLP), and LLMs, all under the Langchain framework. This solution is applicable to both annual and internal audits, enabling real-time verification of compliance data and significantly increasing the speed, accuracy, and reliability of audit procedures, and the proposed workflow is shown in Figure 4.

A. Automated intelligence for yearly auditing document verification

This solution streamlines the annual auditing process by processing documents of varying types, such as images,

scanned documents, and PDFs from users. The text data is extracted using state-of-the-art OCR technologies, i.e., Microsoft's Transformer-based Optical Character Recognition (TrOCR) for hand-written texts, and machine learning-based PDF extraction algorithms. The extracted text is parsed into consumable bits, which are then analyzed using Natural Language Processing (NLP) models, i.e., BERT and RoBERTa, for compliance classification. The Langchain framework is used to handle orchestration of these models, prompt engineering, and RAG support; it comprises retrieving related compliance regulations and contextual information from a vector database and dynamically incorporating them into the generative process, thus enhancing the accuracy and readability of compliance assessments.

Discrepancies and non-compliant areas are highlighted automatically in the document. The solution provides a downloadable, highlighted PDF that visually identifies areas of concern, making the audit review process easier for the auditors. The process not only reduces manual effort but also provides a repeatable, standardized, and auditable compliance process. By automating the abstraction of large volumes of documents, the solution dramatically reduces the turnaround time for year-end audits, eliminates the risk of human error, and provides consistent, high-quality audit output

B. Real-time data extraction and compliance verification during internal audit procedures

In the internal and concurrent audit environment, the system constantly consumes real-time streams of information (e.g. CSVs, PDFs). NLP and OCR pipelines extract text and analyze the extraction in real-time, while Langchain governs pulling relevant compliance standards and coordinates LLM prompts to check for compliance. Compliance is checked against updated regulations, and all results such as compliance milestones, false positives, and overlooked issues are recorded in a single dashboard. If the Compliance Audit Score (CAS) drops below a predetermined threshold, automated alerts are sent to facilitate real-time corrective action. The integration empowers auditors to track compliance status in real-time across various documents and data sources in near real-time. Automation of routine validation activity frees up auditors to concentrate on strategic oversight, anomaly analysis, and value-added analysis. By avoiding delays associated with manual data processing, the system improves the accuracy and efficiency of the internal audit process and offers an end-to-end and current audit trail.

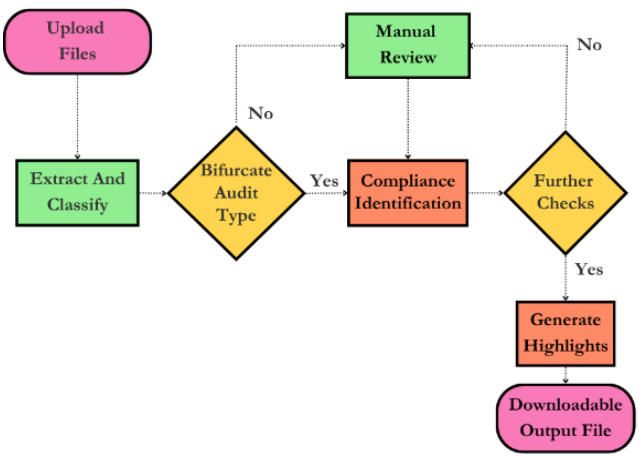

Figure 4. Workflow diagram

5. METHODOLOGY

The proposed methodology introduces a holistic approach to automating compliance auditing through advanced natural language processing and machine learning techniques. The methodology starts with adaptive data ingestion that takes inputs in forms like images, scanned documents, and PDFs.

OCR technology translates every other non-text content into a machine-readable text by utilizing cutting-edge models like Facebook's BART Large MNLI and Microsoft's TrOCR. This system then applies smart text segmentation by the Llama 2 model and Langchain approach to scrape data in a dynamic way based on the requirements of audits.

Langchain acts as the mother model layer for our system, performing several key functions like orchestration of various functions and procedures in a orderly fashion. Langchain is responsible for chunking documents by splitting big documents into semantically relevant pieces of 500-1000 tokens, supporting the processing of long financial documents in an efficient manner. Retrieval Augmented Generation (RAG) is also a possible approach denoted by the framework in the form of having a embedded-vector database containing the set of compliance rules and regulations provided by the authorities, which then goes through a systematic flow of rule matching against document text. Langchain is also responsible for handling prompt engineering and model chaining, with OCR model outputs being systematically input to NLP models using carefully designed prompts that incorporate compliance context and discrete regulatory requirements.

As can be seen from Figure 5, the hybrid method towards extraction and verification of compliance constitutes the research methodology through the emphasis placed on the integration of rule-based methods with deep learning models so as to enable both accuracy in terms of preset regulatory standards and flexibility based on new and changing situations of compliance.

Every model within our pipeline has a specific role. Microsoft's TrOCR addresses handwritten and scanned document text extraction, which is best suited for bank statements and manual entries. Facebook's BART-large-mnli conducts zero-shot classification to assign compliance violations into predetermined classes like "GST Mismatch," "Missing Documentation," or "Regulatory Non-compliance." BERT-base-uncased produces contextual embeddings for semantic similarity matching between document content and regulatory requirements, whereas RoBERTa-base delivers strong text classification for ultimate compliance determination.

Taking the case of a scanned GST invoice uploaded for compliance checking, it first employs TrOCR to extract text from the scanned document, recognizing fields such as "GSTIN: 27AABCU9603R1ZN" and "Tax Amount: ₹18,000." Langchain next chunks this data and fetches applicable GST compliance rules from the knowledge store. BART-large-mnli determines if the GSTIN format is in regulatory compliance, and BERT embeddings determine if the tax computation compares to the expected 18% GST rate for the product type. In case there is a mismatch (e.g., 12% tax charged instead of 18%), the system reports this as "GST_RATE_VIOLATION" and marks the concerned section in the output PDF with a suggestion: "GST rate should be 18% for this product category according to HSN code 8517."

For yearly audits, the system examines the retrieved text for

non-conformity and delivers a holistic output in the form of PDF highlighted areas of nonconformity and recommended remedial measures. For internal audits and concurrent, the system compiles compliance data drawn from various files into one CSV format and analyzes the data to reveal signals of possible risk and recommend measures for remediation of any conformity violations found, following the technological stack as given in Figure 6.

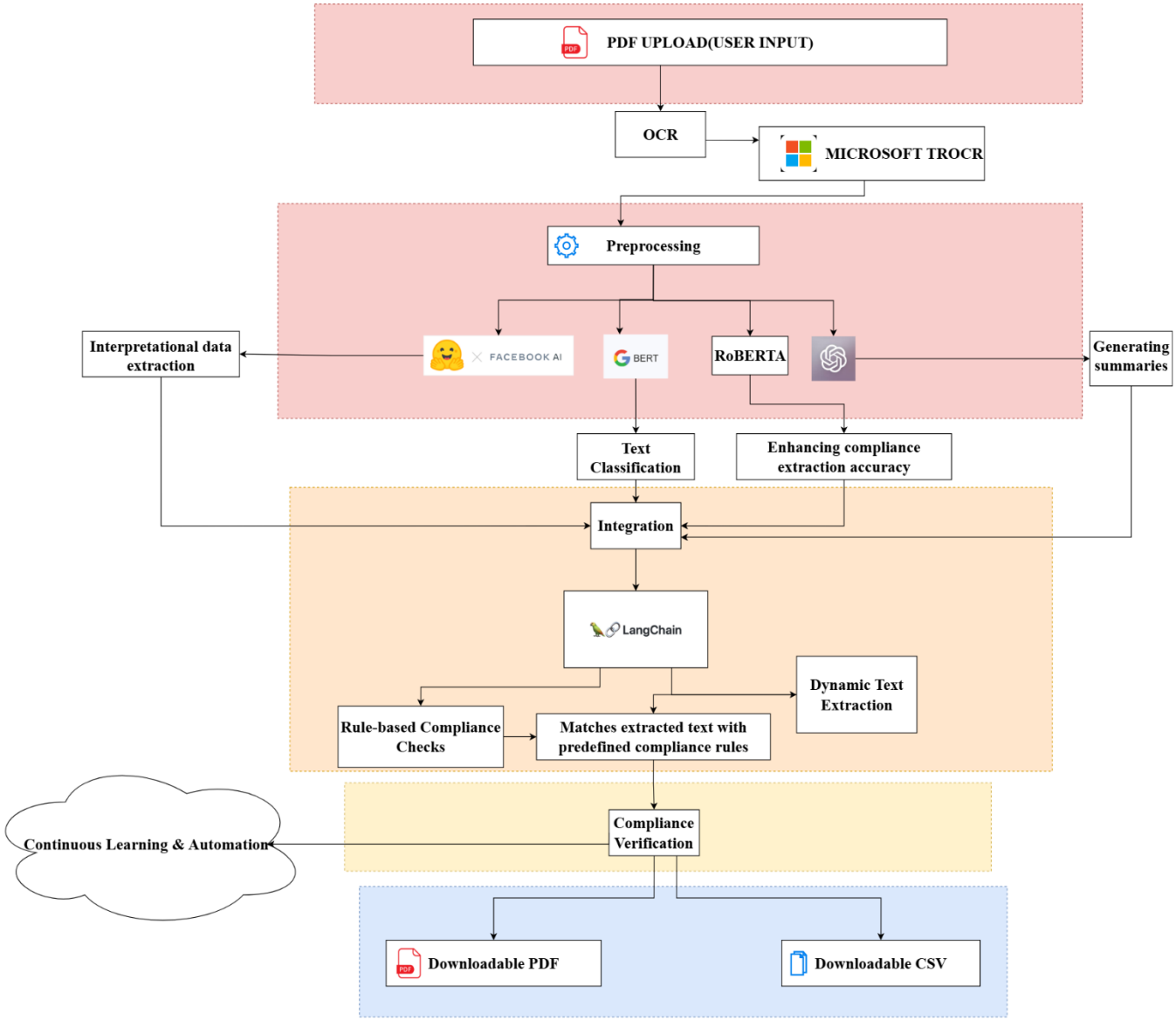

Figure 5. System architecture

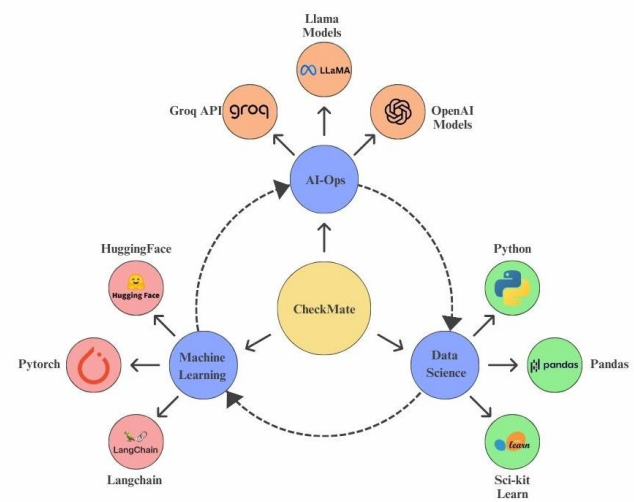

Figure 6. Technical knowledge graph

5.1 Predefined equations of each model

These mathematical equations are instantiated in our system via concrete implementation plans. Every equation is the central computational process that generates the corresponding model's decision-making activity in the pipeline of compliance.

(1) Tesseract (OCR)

Tesseract is based on pattern recognition and machine learning techniques. The primary equation that governs the character recognition process can be represented as:

$$P(C|I) = \frac{P(Ic) * P(c)}{(P(I))} \quad (1)$$

Here,

$P(C|I)$ = Probability of character c given image I

$P(Ic)$ = Likelihood of image I given character c
$P(c)$ = Prior probability of character c
$P(I)$ = Total probability of image I

This Bayesian inference model is used when dealing with poor-quality scanned documents. The system keeps character frequency distributions $P(c)$ of financial document corpora so that financial vocabulary and numerical data often present in audit documents can be better recognized.

(2) Microsoft's TrOCR

TrOCR is designed for text recognition in documents. It uses a transformer architecture for OCR tasks. The foundational equation for a transformer model can be described as following:

$$Z = softmax \tag{2}$$

Here,
Q = Query matrix
K = Key matrix
V = Value matrix
dk = Dimension of the keys

In our methodology, this attention is directed towards financial document layouts, attempting and intending to prioritize numeric values, dates, and regulatory identifiers (such as GST numbers) of essential compliance verification.

a. Bidirectional Encoder Representations from Transformers (BERT)

BERT uses masked language modeling and can be represented as:

$$P(w_i \vee W_1, W_2, \ldots, W_{i-1}, W_{i+1}, \ldots, W_n) = softmax(W_i) \tag{3}$$

Here,
w_1 = Target word
h_1 = Hidden state of the transformer at position i
W = Weight matrix for the output layer

BERT's contextual awareness is used to recognize financial technicalities and regulatory lingo. As it processes terms such as "statutory compliance" or "regulatory compliance," BERT's embeddings recognize the semantic relationships important for proper compliance classification.

b. Bidirectional and Auto-Regressive Transformer (BART)

BART combines bidirectional and autoregressive transformers. The loss function for BART can be defined as:

$$L = -\sum_{t=1}^{T} logP(w_t \vee w \leq t, x) \tag{4}$$

Here,
L = Loss
w_t = Target word at time t
x = Input sequence

BART's sequence-to-sequence feature provides compliance summaries and explanations of violations. Upon detecting a GST mismatch, BART produces readable explanations such as "Expected GST rate of 18% but got 12% - possible under-reporting of tax liability.

c. GPT-4

The core function of GPT-4 is based on autoregressive modeling, expressed as:

$$P(w_T \vee w < t) = softmax(W.h_t) \tag{5}$$

Here,
w_t = Next word in the sequence
h_t = Hidden state of the model at time t

GPT-4 produces context prompts for other models and builds in-depth compliance reports. It compiles the output of several models to create thorough audit summaries with actionable advice.

5.2 Merged compliance assessment formula

The CAS can be formulated to capture the contributions of both OCR and NLP models in the compliance checking process.

$$CAS = \alpha * \sqrt{(POCRxPNLP)} + \beta * (\frac{TP^2}{TP + FP + FN}) \tag{6}$$

Here,
$POCR$ = Effectiveness of the OCR component (probability of correct character recognition)
$PNLP$ = Effectiveness of the NLP component (accuracy of language models in identifying compliance)
TP = True Positives (correctly identified compliant items)
FP = False Positives (incorrectly flagged items)
FN = False Negatives (missed compliant items)
α and β are weighting coefficients to balance the contributions of the OCR and NLP components.

This weighted score is computed in real-time while documents are processed. The empirically set weighting coefficients α (0.4) and β (0.6) reflect the relative significance of OCR accuracy vs. NLP comprehension in compliance detection. For example, in case of audits involving handwritten documents, α is raised to 0.6 as a result of increased OCR reliance. When a batch of 100 invoices are processed, and the system identifies 85 correct compliant documents (TP = 85), incorrectly identifies 5 as non-compliant (FP = 5), and fails to identify 3 actual violations (FN = 3), with OCR accuracy of 95% and NLP accuracy of 96%, then CAS will be calculated as: CAS = 0.4 × (85/90) × 0.95 + 0.6 × (85/88) × 0.96 = 0.357 + 0.558 = 0.915, reflecting high system performance.

Algorithm 1. Automated intelligence for yearly auditing verification

Input: Document (various formats)
Output: CAS and marked-up document

```
Initialize TP, FP, FN to 0
Use OCR to extract text from the document
Segment the extracted text into manageable chunks
for each compliance rule in the set:
    Check if the rule applies to the text chunk
    If compliant:
        Increment TP
    Else if flagged incorrectly:
        Increment FP
    Else:
```

Increment FN
Calculate Precision using the formula defined above
Calculate Recall using the formula defined above
Calculate OCR Effectiveness using the formula defined above
Calculate NLP Effectiveness using the formula defined above
Calculate Compliance Audit Score (CAS) using the formula defined above
Generate highlighted document with compliance results (marked-up PDF)

Algorithm 2. Extracting real time data and verifying compliance

Input: Live data stream (various formats: PDFs, CSVs)
Output: CAS and marked-up document

Initialize TP, FP, FN to 0
While documents are being processed:
    Extract text from the live document using OCR
    Segment the extracted text into manageable chunks
    For each compliance rule in the set:
        Check if the rule applies to the text chunk
        If compliant:
            Increment TP
            Log compliance success
        Else if flagged incorrectly:
            Increment FP
            Log false positive
        Else:
            Increment FN
            Log missed compliance
    Calculate Precision and Recall using the formula defined above
    Calculate OCR, NLP and CAS Effectiveness using the formula defined above
    If CAS < threshold:
        Trigger alert for compliance issues

6. RESULTS

Our research on AI-driven financial auditing and compliance systems yielded several significant findings, encompassing safeguards for AI implementation, performance of various AI models, and compliance checks for five major Indian banks. The survey of industry experts revealed critical safeguards necessary for implementing AI in financial auditing and compliance. Figure 7 illustrates the prioritization of these safeguards.

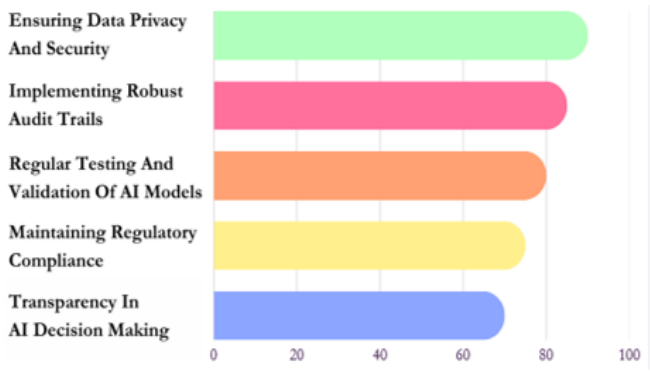

Figure 7. Safeguard measures for AI in financial auditing and compliance

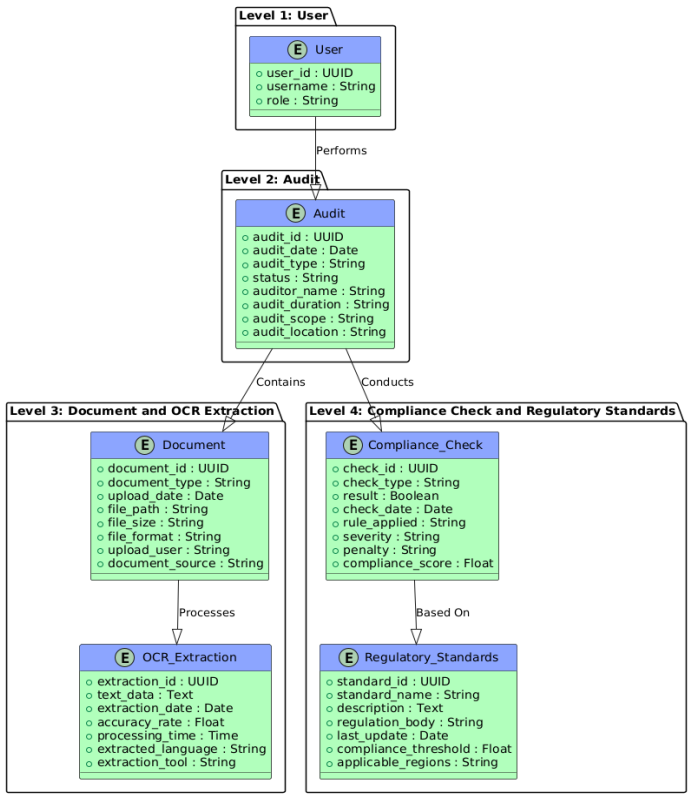

Figure 8. E-R diagram

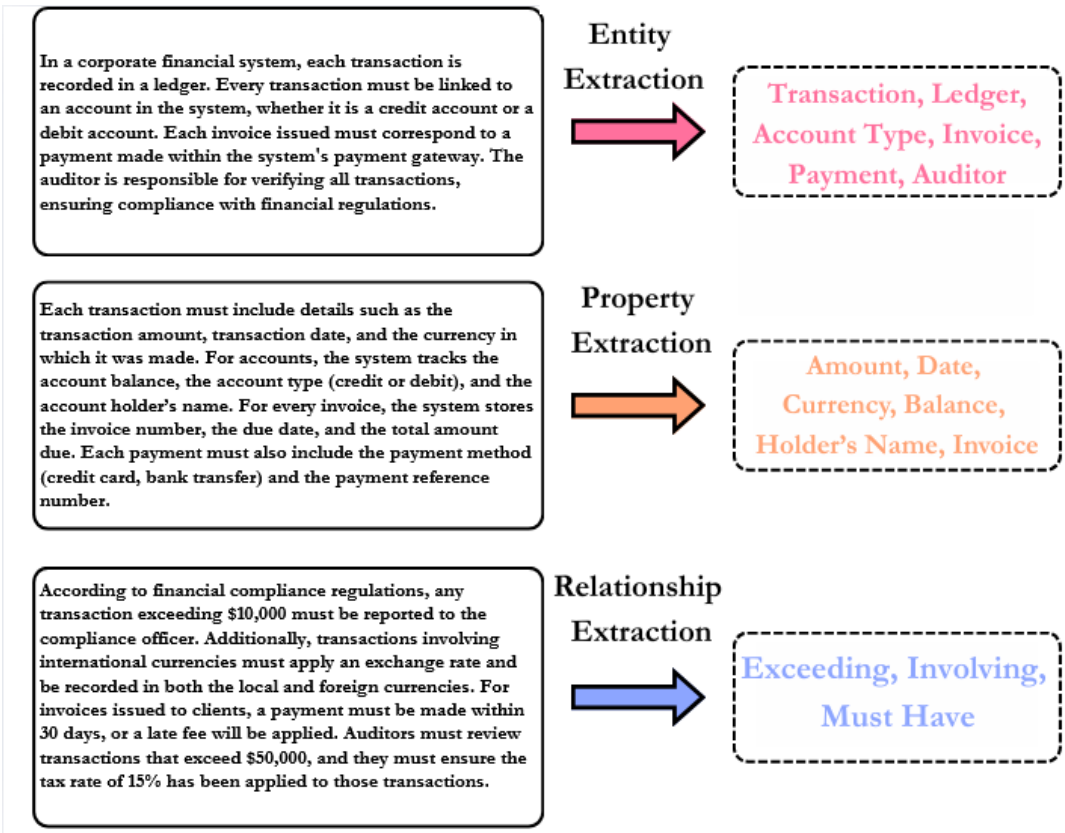

Figure 9. Extraction and highlights

The results highlight the paramount importance of data privacy and security, with 90% of experts emphasizing this safeguard. Implementing robust audit trails (85%) and regular testing and validation of AI models (80%) were also highly prioritized. Maintaining regulation, transparency of how AI has resolved a specific problem, and yet still under human observation obtained critical ranks as 75%, 70%, and 65%. These findings direct and amplify the need for a comprehensive wholistic approach to automation software's implementation in financial auditing and compliance checking processes, following the schema from Figure 8.

This is how the highlighted downloadable functionality of data conversion, isolation, chunking and hence execution of the checking function on the PDF would work, using NLP techniques powered by Langchain on the extracted text chunks splitted as in an Augmented Retrieval System, further to be used with embedded vector databases, refer to Figure 9 for the same.

Evaluation of AI models for various tasks related to financial auditing and checking compliance, we observed notable performance differences. The following Table 1 summarizes the comparison results:

Table 1. Comparative results

Model Name	Accuracy	Input Type	Output Type
Facebook/Bart-large-mnli	96.8%	Text	Classification
Microsoft/TROCR-base-handwritten	95.2%	Image	Text
Bert-base-uncased	94.7%	Text	Embeddings
GPT2	96.3%	Text Prompt	Generated Text
Roberta-base	95.9%	Text	Classification

The Facebook/BART-large-mnli model achieved the highest accuracy (96.8%) for compliance checking tasks, while the Microsoft/TROCR-base-handwritten model excelled in OCR for handwritten text with 95.2% accuracy. For text classification, both Bert-base-uncased and Roberta-base showed high accuracy (94.7% and 95.9% respectively). The GPT2 model, while slightly less accurate (96.3%), demonstrated superior performance in text generation tasks. The results are probably that the multiple models strategy would deliver best solutions to real applications of AI in finance-based auditing and compliance.

The following Table 2 summarizes the KPIs of automation auditing (proposed solution) and manual auditing and its resultant metrics as follows:

Table 2. KPI results

Metric	Manual Auditing	Automated Auditing
F1-score	0.82	0.95
CAS threshold	N/A	0.90
Accuracy in identifying non-compliance	85%	95%
Pages processed per hour	50-100	1,000
Average audit completion time	4 weeks	2 weeks
Human error rate	~15%	~6%
Estimated 3-year cost savings	$0	$2 million*

Note: *Assumes a large bank with >$50 billion in assets, 50% reduction in audit time, and proportional reduction in manpower costs.

Table 2 provides a comparison of automated and manual auditing procedures, and we can clearly see the enhanced efficiencies and accuracy with the proposed system that will occur with the implementation of automation. Although

manual systems handle 50-100 pages/hour with ~15% error, spend approximately four weeks on an audit [10, 12, 13], the proposed system handles +1,000 pages/hour, reduces errors to 6%, and completes an audit in two weeks. Its F1-score is 0.95 with a CAS of 0.90, and the automated system provided more accurate audits. We have tested the practical applicability of our AI-driven auditing system by conducting a deep compliance check of five major Indian banks, as indicated in Figure 10.

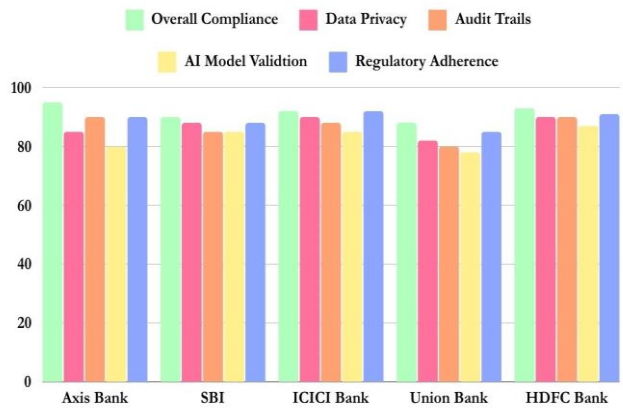

Figure 10. Overall compliance results for Indian banks

Our AI-driven compliance check reflects that the five major Indian banks have levels of the requirements of the regulatory body, and the dashboard offers such metrics as overall compliance score, data privacy measures, audit trail robustness, and AI model validation processes about regulatory adherence.

These results justify the approach and showcases the benefits of our AI-driven approach towards the accuracy, efficiency, and holistic development in the sector of financial auditing and compliance-check methods in Indian banks. The dashboard would hence visually be able to deliver what is strong and what needs improvement within a bank's system of the trails and their management so that interventions have that targeted, precise and directed strategic intent. The high-performance software system developed on our research also holds a dimension support enhancement of such compliance measures and their operational integrity in these procedures followed by the banks in India.

Even the differences in evaluation metrics along the broad spectrum of compliance aspects thus point towards the complexity and the shortcomings of regulatory adhesion in the existing practices of the banking industry.

7. CONCLUSION AND FUTURE WORK

7.1 Conclusion

As the attempt was intended, an implementation of such system that holds a capability of turning automated compliance checking procedures a reality, can be significantly useful to thereby optimize the resources and fortify outcomes through a significant level, creating impact by also cutting down overall time. The results showcase high quantifiable value along all the possible dimensions, with a very few downsides that can be tackled well in the future with advancing technologies. For instance, the 75% reduction of manual document review time enables the audit teams to save an average time of 20 hours per week. Almost a 10x productivity increase, from the traditional 50–100 pages compared to 1,000 pages an hour. The system also holds the capability of reducing human error by enlarge, highlighting the flaws in a precise yet cited format. From a financial standpoint, mid-sized banks can expect to reduce audit costs by 40% over 3–4 years. Adapting such software solutions aim compliment the integrity and enhance transparency overall amid the process.

7.2 Limitations

a. A challenge in such an AI-model led approach is to officially gather huge chunks of real-world data in order to train the model effectively on a domain, be it any type of compliance, which is why the current evaluations are limited to Indian Banks, performed through training on a small set of publicly available annual audits data from the official websites, being entirely in legal bounds, enough datasets should be used to train a production-level model, only then decent accuracy scores can be expected, avoiding all kinds of model drifts if the data is not sufficient, covering all kinds of edge-cases.

b. While such an approach can yield optimized outcomes overall, an area of concern here is all of the audit data, be it scanned copies, PDFs, or even the regulations, along with the satisfied or unsatisfied results are logged on to the AI-model that fuels the entire system, such an event can cause a fatal exposure of data to be leaked or breached, to avoid such circumstances, strong encryption algorithms, like AES and CBC-paddings should be enforced, so that no direct transmission of raw data occurs, compromising the overall integrity of the process.

7.3 Future discussion

This research is the basis of a promising possibility, laying a strong as well as reasonable foundation for a progressive shift towards an automated financial auditing and compliance checking system. Future developmental aspects can definitely cover developing a system for real-time support with comprehensive real-time regulatory updates and implementation of blockchain technology to significantly enhance the security measures, blockchain ledger could be distributed among trusted parties, instead of a singular internal database, to enhance security aspects, also once written, the record of the audit log can't be altered without leaving a trace, any modification would be cryptographically evident at any point in the process, such innovations can direct us to federated learning possibilities incorporated within the system for auditing and checking compliance. In addition, even Retrieval Augmented Generation (RAG) models can be adapted with regards to the general government guidelines for banks, and correspondingly notices as per the authority and as per company officers for private institutions, to keep the overall context of the narrative better in-control, in turn fetching better results. These developments would complement the system to be an epicenter of advanced financial auditing.

REFERENCES

[1] Arner, D.W., Barberis, J., Buckley, R.P. (2016). The evolution of fintech: A new post-crisis paradigm. Georgetown Journal of International Law, 47: 1271-1319. http://doi.org/10.2139/ssrn.2676553

[2] Lee, I., Shin, Y.J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. Business Horizons, 61(1): 35-46. https://doi.org/10.1016/j.bushor.2017.09.003

[3] Milian, E.Z., Spinola, M.M., Carvalho, M.M. (2019). Fintechs: A literature review and research agenda. Electronic Commerce Research and Applications, 34: 100833. https://doi.org/10.1016/j.elerap.2019.100838

[4] Schueffel, P. (2016). Taming the beast: A scientific definition of fintech. Journal of Innovation Management, 4(4): 32-54. https://doi.org/10.2139/ssrn.3097312

[5] Butler, T., O'Brien, L. (2019). Understanding RegTech for digital regulatory compliance. In Disrupting Finance, Palgrave Pivot, Cham, pp. 85-102. https://doi.org/10.1007/978-3-030-02330-0_6

[6] Power, M. (2013). The apparatus of fraud risk. Accounting, Organizations and Society, 38(6-7): 525-543. https://doi.org/10.1016/j.aos.2012.07.004

[7] Rezaee, Z., Sharbatoghlie, A., Elam, R., McMickle, P.L. (2002). Continuous auditing: Building automated auditing capability. Auditing: A Journal of Practice and Theory, 21(1): 147-163. https://doi.org/10.2308/aud.2002.21.1.147

[8] Appelbaum, D., Kogan, A., Vasarhelyi, M.A. (2017). Big Data and analytics in the modern audit engagement: Research needs. Auditing: A Journal of Practice and Theory, 36(4): 1-27. https://doi.org/10.2308/ajpt-51684

[9] Gepp, A., Linnenluecke, M.K., O'Neill, T.J., Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. Journal of Accounting Literature, 40: 102-115. https://doi.org/10.1016/j.acclit.2017.05.003

[10] Li, H., Dai, J., Cheng, T., Vasarhelyi, M.A. (2020). On the application of robotic process automation in banking industry. Journal of Emerging Technologies in Accounting, 17(2): 71-88. https://doi.org/10.2308/JETA-19-04-02-42

[11] Cooper, L.A., Lowe, D.J., Bierstaker, J.L. (2021). The effects of artificial intelligence on audit efficiency and audit quality. Journal of Emerging Technologies in Accounting, 18(1): 31-47. https://doi.org/10.2308/JETA-19-04-21-45

[12] Sun, T., Vasarhelyi, M.A. (2017). Deep learning and the future of auditing: How an evolving technology could transform analysis and improve judgment. CPA Journal, 87(6): 24-29. https://www.cpajournal.com/2017/06/19/deep-learning-future-auditing/.

[13] Verma, S., Jain, K., Prakash, C. (2020). An unstructured to structured data conversion using machine learning algorithm in Internet of Things (IoT). SSRN Electronic Journal. http://doi.org/10.2139/ssrn.3563389

[14] Eilifsen, A., Kinserdal, F., Messier, W.F., McKee, T.E. (2020). An exploratory study into the use of audit data analytics on audit engagements. Accounting Horizons, 34(4): 75-103. https://doi.org/10.2308/HORIZONS-19-121

[15] Hou, B., Zhang, Y., Shang, Y., Liang, X. (2020). Research on unstructured data processing technology in executing audit based on big data budget. Journal of Physics: Conference Series, 1684(1): 012014. https://doi.org/10.1088/1742-6596/1650/3/032100

[16] Kanimozhi, K.V., Venkatesan, M. (2015). A review on unstructured data management. International Journal of Advanced Research in Computer and Communication Engineering, 4(3): 188-191. https://doi.org/10.17148/IJARCCE.2015.4354

[17] Reddy, R.V.K., Venugopal, G., Rajanala, G., Sambasivarao, V., Harshavardhan, N., Ajay, T. (2020). Transforming unstructured data to structured data using map reduce and HBase. International Journal of Emerging Trends in Engineering Research, 8(9): 6117-6121. https://doi.org/10.30534/ijeter/2020/137892020

[18] Hansen, M., Pomp, A., Erki, K., Meisen, T. (2019). Data-driven recognition and extraction of PDF document elements. Technologies, 7(3): 65. https://doi.org/10.3390/technologies7030065

[19] Ashish. (2024). Data extraction from unstructured PDFs. https://www.analyticsvidhya.com/blog/2021/06/data-extraction-from-unstructured-pdfs/.

[20] Li, H., Gao, H., Wu, C., Vasarhelyi, M.A. (2023). Extracting financial data from unstructured sources: Leveraging large language models. Journal of Information Systems, 39(1): 135-156. http://doi.org/10.2139/ssrn.4567607

[21] Sarmah, B., Mehta, D., Pasquali, S., Zhu, T. (2023). Towards reducing hallucination in extracting information from financial reports using large language models. International Conference on AI-ML-Systems. https://doi.org/10.48550/arXiv.2310.10760

[22] Liu, X.Y., Wang, G., Yang, H., Zha, D. (2023). Fingpt: Democratizing internet-scale data for financial large language models. arXiv preprint arXiv:2307.10485. https://doi.org/10.48550/arXiv.2307.10485

[23] Nam, B., Rawte, V., Zaki, M.J., Gupta, A. (2023). FETILDA: An effective framework for fin-tuned embeddings for long financial text documents. arXiv preprint arXiv:2206.06952. https://doi.org/10.48550/arXiv.2206.06952

[24] Zhao, H., Liu, Z., Wu, Z., Li, Y., et al. (2024). Revolutionizing finance with llms: An overview of applications and insights. arXiv preprint arXiv:2401.11641. https://doi.org/10.48550/arXiv.2401.11641

[25] Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., Staiano, J. (2023). Glitter or gold? Deriving structured insights from sustainability reports via large language models. EPJ Data Science, 13(1): 1-41. https://doi.org/10.1140/epjds/s13688-024-00481-2

[26] Prescott, E.S. (2004). Auditing and bank capital regulation. FRB Richmond Economic Quarterly, 90(4): 47-63. https://ssrn.com/abstract=2184962.