Using Large Language Models to Support the Audit Process in the Accountability of Interim Managers in Notary Offices

Myke Valadão[a], Natalia Freire[b], Mateus de Paula[c], Lucas Almeida[d] and Leonardo Marques[e]
SiDi Intelligence & Innovation Center, Manaus, Brazil
{*m.douglas, natalia.f, mp.silva, l.dasilva, lc.marques*}*@sidi.org.br*

Keywords: Auditing, Large Language Models, Fraud Detection, Notary Offices, Resource Management, AI.

Abstract: The auditing process in notary offices in Brazil is hindered by inefficiencies, high costs, and the complexity of manual procedures. To address these challenges, we propose a system that leverages the capabilities of Large Language Models (LLMs), specifically LLaMA2-7B and Falcon-7B, to automate critical information extraction from diverse document types. The system detects anomalous monetary values and unauthorized services, linking them to corresponding dates and beneficiaries to provide a detailed overview of financial discrepancies. Integrating advanced Natural Language Processing (NLP) techniques into auditing workflows enhances fraud detection, reduces operational costs, and improves accuracy. With a BLEU metric superior to 0.67, the proposed system demonstrates significant potential to streamline auditing operations. Key benefits include assisting court analysts in identifying fraud cases, optimizing public resource management by eliminating unjustified expenses, and potentially increasing court revenues to reinvest in public services.

1 INTRODUCTION

The audit process in notary offices presents challenges that significantly impact efficiency, financial sustainability, and accountability. One of the primary issues is the high financial cost associated with manual auditing procedures. Auditors must manually examine and cross-reference large volumes of documents, which is both time-consuming and labor-intensive. Furthermore, the diversity of document types, including contracts, deeds, invoices, and receipts, adds complexity to the process. Each document type may have a unique structure, format, and set of relevant data points, making it challenging to apply a standardized approach. Additionally, these manual processes make it harder to detect and prevent fraudulent activities, such as falsified reports, superfluous charges, and unauthorized or prohibited services (Santana et al., 2024; Simunic, 1980). This reliance on manual methods increases the likelihood of human error and limits the scalability of auditing operations, ultimately undermining the transparency and effectiveness of interim managers in notary offices.

To address these issues, integrating LLMs offers a transformative solution. With their advanced NLP capabilities, LLMs can automate and streamline the extraction and analysis of critical information from diverse document types. By doing so, they can significantly reduce the time and cost associated with manual audits. Moreover, LLMs are adept at identifying anomalies, such as inconsistencies in financial data or suspicious activities, which can aid in fraud detection and prevention. These models also ensure greater accuracy and consistency in data processing, enabling interim managers to generate more reliable reports and make informed decisions. Ultimately, using LLMs enhances the overall audit process, improving efficiency and accountability in notary offices.

LLMs are advanced neural networks designed to understand, generate, and analyze human language (Karanikolas et al., 2023). They are trained on vast amounts of text data, ranging from books and articles to legal and financial documents (Kumar, 2024). Through this training, LLMs learn to recognize patterns, relationships, and contextual nuances in language, enabling them to perform various tasks such as text summarization, information extraction, and anomaly detection. Examples of widely known LLMs include OpenAI's GPT series, Meta's LLaMA, and Google's Gemini. These models are particularly valu-

[a] https://orcid.org/0000-0001-7595-2266
[b] https://orcid.org/0000-0002-0762-9800
[c] https://orcid.org/0009-0009-9060-6447
[d] https://orcid.org/0009-0002-7106-248X
[e] https://orcid.org/0000-0002-3645-7606

Valadão, M., Freire, N., de Paula, M., Almeida, L. and Marques, L.
Using Large Language Models to Support the Audit Process in the Accountability of Interim Managers in Notary Offices.
DOI: 10.5220/0013480900003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 988-995
ISBN: 978-989-758-749-8; ISSN: 2184-4992

able in notary office settings, where they can rapidly process large datasets with high precision. LLMs enable real-time auditing and reporting by extracting key information from complex documents, such as dates, monetary values, and descriptions. Additionally, they assist in generating summaries, translating legal terminology, and flagging potential discrepancies, thus providing comprehensive support throughout the audit process.

Building on these capabilities, we propose a system leveraging the LLaMA2-7B and Falcon-7B models to automate the critical information extraction from notary office documents. This system is designed to identify anomalous monetary values and unauthorized services by analyzing various printed documents. Once anomalies are detected, the system associates these irregularities with the corresponding dates and beneficiaries, providing a comprehensive overview of financial discrepancies. By integrating these LLMs into the auditing workflow, our proposal aims to enhance the detection of fraudulent activities and streamline the accountability of interim managers in notary offices, ultimately improving transparency and operational efficiency. The proposed method achieved a BLEU metric superior to 0.67, demonstrating the system's effectiveness in accurately extracting critical financial information and its potential to enhance the auditing process in notary offices. Additionally, The system enhances auditing by automating financial analysis, enabling analysts to focus on fraud detection, optimizing public resource allocation, and increasing judicial revenue for reinvestment in public services.

The remainder of this paper is organized as follows: Section 2 presents the fundamental concepts necessary for understanding this research, along with related work. Section 3 outlines the AI methods employed to support the audit process discussed in this paper. Section 4 details the experiments we carried out, and the results obtained from the AI methods. Finally, Section 5 concludes the paper with final remarks and directions for future work.

2 BACKGROUND AND RELATED WORK

2.1 Notary Offices and Managers

Extrajudicial notary offices play a vital role in the legal system, handling civil registration, deeds, real estate documentation, and signature notarization. These services are provided by tenured notaries, but vacancies may arise due to unforeseen events, requiring a temporary appointment to ensure continuity. The interim notary, designated by a competent authority such as the General Inspectorate of Justice or a district judge, assumes responsibility for managing the office during the vacancy.

The interim officer's duties include overseeing administrative operations, executing notarial acts, ensuring compliance with legal norms, reporting financial accountability, and fulfilling judicial orders. Our focus is specifically on accountability, where we explore how LLMs can assist court analysts in detecting anomalies in financial reports submitted by interim officers. By leveraging LLM-based anomaly detection, we aim to enhance oversight efficiency and accuracy.

2.2 Related Work

Devlin (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), a powerful transformer-based model that revolutionized the field of NLP. BERT's architecture leverages bidirectional context, enabling it to understand a word's meaning based on its preceding and following context. This approach has been successfully applied to various NLP tasks, including document classification and information extraction. It is highly relevant to auditing processes where extracting specific data points from diverse document types is crucial. Despite its impressive performance, BERT's application in financial auditing and anomaly detection remains underexplored, particularly in adapting the model for domain-specific challenges like notary office document analysis.

Rudžionis et al. (2022) presents a sophisticated approach to detecting irregular financial transactions by applying NLP techniques. The researchers focus on accountant comments within ledger entries, often containing informal yet insightful annotations. By leveraging cosine similarity and term frequency-inverse document frequency (TF-IDF) methods, they analyze the semantic content of these comments to identify outlier financial operations. Their experiments, conducted on data from Dutch companies, revealed that only 0.3% of transactions flagged for further review were potentially suspicious, substantially reducing the workload for human auditors. However, the study notes that the model's performance may vary depending on the quality and availability of comments, highlighting a need for further refinement and application across diverse datasets.

Beltran (2023) explores how NLP can enhance the extraction of critical fiscal data from audit reports, particularly in the context of subnational supreme au-

dit institutions (SAIs) in Mexico. The research focuses on audit reports from Sinaloa's SAI, systematically converting scanned documents into machine-readable text using Optical Character Recognition (OCR). It then employs a text classification model to filter relevant content and a named entity recognition (NER) system to extract monetary values associated with budget discrepancies. The paper underscores the importance of leveraging machine learning for large-scale document analysis, addressing inefficiencies and accessibility issues in traditional auditing. While the approach is promising, potential areas for further research include expanding the methodology to audits from different regions and improving the handling of OCR errors for non-English languages.

Fisher et al. (2016) provide an extensive review of NLP applications in these domains, highlighting its use for classification, information retrieval, fraud detection, and financial predictions. The authors emphasize the interdisciplinary nature of NLP and its potential to transform traditional financial and audit processes. They discuss various computational methods to extract insights from unstructured textual data, including text mining, sentiment analysis, and machine learning. Despite progress, the paper identifies gaps in the literature, such as the need for more refined tools for taxonomy generation and integrating NLP with continuous auditing systems. These areas present opportunities for further exploration, particularly in enhancing the automation and accuracy of financial document analysis.

3 METHODS

This section describes the approach to addressing the use of LLMs to extract crucial information from documents to assist interim managers in notary offices.

3.1 System Model

The role of interim managers in notary offices is complex, requiring them to process large volumes of legal, administrative, and financial documents. Manual data extraction from these documents is time-consuming and error-prone, particularly when dealing with unstructured or semi-structured formats like contracts, deeds, and invoices. Ensuring accuracy and consistency is critical, as errors may lead to legal or financial discrepancies. Additionally, time-sensitive deadlines demand an efficient system for handling large-scale document processing.

To address these challenges, we propose leveraging LLMs for automated information extraction from printed document images. Advanced NLP models like GPT and BERT (Zheng et al., 2021) can efficiently process structured text, mitigating the complexities of handwriting recognition, which suffers from high variability, poor resolution, and segmentation issues (Alhamad et al., 2024; Ingle et al., 2019). By focusing on printed documents, we enhance accuracy and efficiency, enabling the identification of outlier values along with relevant dates, descriptions, and creditors.

As illustrated in Figure 1, our system consists of four key components: preprocessing, OCR, LLM-based extraction, and performance evaluation. Preprocessing enhances document images, while OCR converts them into machine-readable text. The LLM then extracts and organizes critical data points, supporting legal and financial reporting. Finally, an evaluation step ensures accuracy and consistency, streamlining document analysis in notary offices.

3.2 Preprocessing

In the preprocessing, the following steps are performed:

1. Binarization: Converts the input image to a binary format, enhancing contrast and eliminating noise for improved OCR accuracy. This step is crucial for unstructured or poorly scanned documents (Saini, 2015).
2. Alignment: Corrects misaligned text by evaluating skew angles θ within a range ℓ. A scoring function $S(\theta)$ determines the best alignment, and the document is rotated accordingly to produce a corrected version I', improving OCR and LLM performance (de Elias et al., 2019).
3. Printed or Handwritten Classification: A CNN-based classifier distinguishes between printed and handwritten text. The model extracts spatial features and assigns a probability p of a document being printed. If $p \geq \tau$, it follows the OCR pipeline; otherwise, it is flagged for specialized processing or manual review (Alhamad et al., 2024)(Ingle et al., 2019).
4. Text Line Detection: The CRAFT model detects and segments text lines by identifying character regions and linking them into coherent structures. Given an input image I', CRAFT generates bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$ representing detected text lines, refining overlapping regions and filtering noise. This ensures accurate text extraction while preserving document structure, making it essential for handling diverse formats and layouts (Baek et al., 2019).

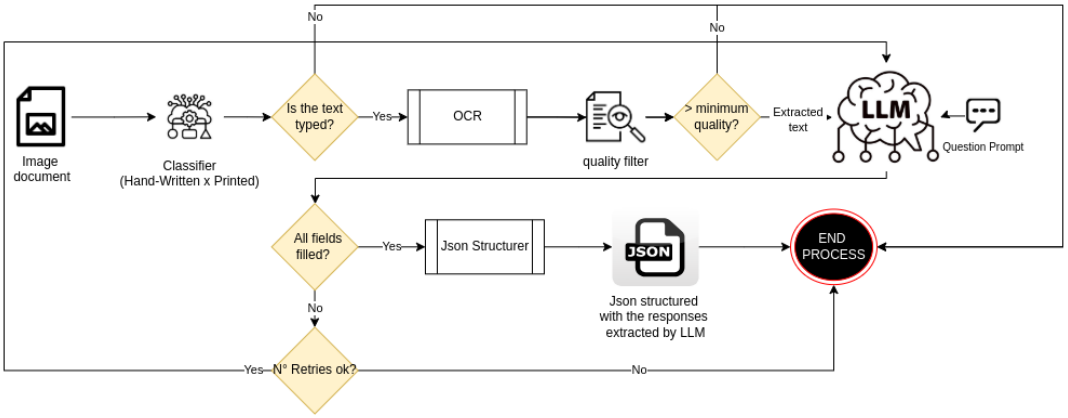

Figure 1: Complete system for extracting valuable information from documents.

3.3 OCR

OCR technology converts images of text into machine-readable text, enabling automated document processing and analysis. Tesseract OCR, an open-source tool developed by Google, is used to extract text from images, employing connected component analysis to identify text regions and segment the image into lines and words. It utilizes pattern matching and statistical modeling, along with language-specific dictionaries, to improve accuracy and correct errors. Tesseract supports multiple languages and scripts and can handle complex layouts, such as tables or multi-column text, making it highly versatile for document digitization (Hegghammer, 2022).

Tesseract works best with preprocessed images, where noise, skew, and distortions are minimized. The CRAFT model, which highlights text regions, improves Tesseract's focus on text, enhancing both speed and accuracy (Hegghammer, 2022). The output can be plain text or structured formats like HOCR or ALTO XML, which preserve the spatial layout and provide additional details such as word coordinates, font size, and confidence scores. This flexibility allows Tesseract to cater to a range of use cases, from simple text extraction to complex document analysis (Hegghammer, 2022).

3.3.1 Filter of Quality

Ensuring the reliability of recognized text is crucial, especially when dealing with noisy or low-resolution documents. To address this, our system implements a confidence-based *Filter of Quality*, leveraging Tesseract OCR's confidence scores, which range from 0 to 100. For each recognized token, Tesseract assigns a confidence value, and the system calculates an average confidence score $\bar{c}$ across all detected tokens in a document segment:

$$\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i \tag{1}$$

where c_i represents the confidence of the i-th token, and N is the total number of tokens. If $\bar{c} \geq 40$, the text is considered reliable for downstream processing, including entity recognition and monetary value extraction. Otherwise, it is flagged as low confidence and either rerouted for manual verification or subjected to additional reprocessing, such as enhanced image preprocessing or alternative OCR models.

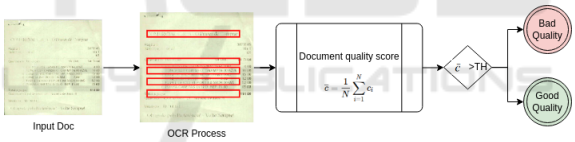

Figure 2: Graphical representation of the confidence-based *Filter of Quality* mechanism. This diagram illustrates the workflow for calculating average confidence scores and determining their suitability for further processing.

As illustrated in Figure 2, this filtering mechanism prevents unreliable text from contaminating later analysis stages, such as anomaly detection or LLM processing. The threshold $\tau = 40$ was determined through empirical tuning, balancing the need to retain useful text while minimizing recognition errors. This approach improves workflow efficiency by allowing a more targeted document auditing process, ensuring that only high-confidence text proceeds for further analysis.

3.4 LLM for Extracting Information

3.4.1 Prompt Engineering

In the context of extracting critical information from documents, the quality of the OCR output plays a pivotal role. However, OCR often produces outputs that

lack clarity and coherence, particularly when dealing with unstructured or semi-structured documents. These outputs can include incomplete sentences, misplaced characters, and inconsistent formatting, which pose challenges for downstream tasks like information extraction. We leverage prompt engineering to optimize the interaction between the extracted text and the LLM to address this (Grabb, 2023).

The designed prompt serves as a structured query framework that enhances the LLM's ability to comprehend and process the noisy OCR output (Wang et al., 2024). Specifically, the prompt includes contextual instructions and specific task directives to guide the model in identifying and extracting key pieces of information. This includes critical details such as dates, monetary values, creditor names, and descriptive content related to transactions or agreements. We improve the model's accuracy and reliability in parsing unclear or fragmented text by framing the prompt to include examples, clarifications, and explicit extraction goals.

Moreover, the prompt dynamically adapts to the input text's structure by instructing the LLM to infer missing details or correct minor inconsistencies. This methodology enhances the precision of extracted data and ensures that essential information is consistently retrieved across diverse document formats (Wang et al., 2024). In doing so, the prompt becomes a critical bridge between the raw OCR output and the structured, actionable insights required by Interim Managers in notary offices.

Example of possible prompt:

```
"You are a highly intelligent assistant
tasked with extracting specific
information from a text document. The
text may contain noise, be incomplete,
or lack formatting. Your goal is to
extract the following details:

Date: (e.g., 2023-05-14, May 14, 2023)
Monetary Value: (e.g., $1,250.00
                 or R$ 1.250,00)
Creditor Name: (e.g., ABC Corporation,
               John Doe)
Description: (e.g., Payment for services,
             Invoice for shipment)
This is the text: {Output of the OCR}"
```

3.4.2 Proposed LLMs

The Falcon-7B and LLaMA2-7B models have been selected for this study due to their advanced architectures and suitability for handling unstructured data. Falcon-7B, developed by the Technology Innovation Institute, leverages a multi-query attention mechanism and pre-normalization to ensure computational efficiency and faster inference (HuggingFace, 2024). In contrast, LLaMA2-7B, designed by Meta AI, incorporates multi-head attention and task-specific optimizations, offering enhanced contextual understanding at the cost of higher memory usage (MetaAI, 2024). Both models utilize rotary positional embeddings (Almazrouei et al., 2023)(Touvron et al., 2023), ensuring stability and adaptability across diverse tasks. Additionally, both models were trained using mixed precision techniques, which allow for efficient utilization of modern graphics processing unit (GPU) resources. This complementary use of Falcon-7B and LLaMA2-7B ensures robust performance in extracting critical information from complex document structures. The key architectural differences between these two models are summarized in Table 1.

Table 1: Comparison of Falcon-7B and LLaMA2-7B Architectures.

Aspect	Falcon-7B	LLaMA2-7B
Transformer Block Design	Standard transformer with RoPE (Almazrouei et al., 2023)	Enhanced transformer with RoPE and task-specific optimizations (Touvron et al., 2023)
Attention Mechanism	Multi-query attention (MQA) for efficiency (Almazrouei et al., 2023)	Grouped-query attention (GQA) with multi-head baseline (Touvron et al., 2023)
LayerNorm Placement	Pre-normalization for stability (Almazrouei et al., 2023)	Pre-normalization for efficiency (Touvron et al., 2023)
Feedforward Network (FFN)	GELU activation (Almazrouei et al., 2023)	SwiGLU activation (Touvron et al., 2023)
Parameter Efficiency	Lower computational cost (Almazrouei et al., 2023)	Task-specific adaptability (Touvron et al., 2023)
Training Precision	Mixed precision (BF16) (Almazrouei et al., 2023)	Mixed precision for GPU compatibility (Touvron et al., 2023)
Inference Optimization	Optimized for memory efficiency (Almazrouei et al., 2023)	Focuses on precision with higher memory usage (Touvron et al., 2023)

3.5 Evaluation

The classification model was evaluated using Precision, Recall, and F1-Score:

- **Precision**: Proportion of correctly predicted positives among all positive predictions:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

- **Recall**: Proportion of correctly predicted positives among actual positive cases:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

- **F1-Score**: Harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

For text similarity evaluation, we employed BLEU and Cosine Similarity:

- **BLEU**: A precision-based metric comparing n-grams of generated and reference texts, with a brevity penalty to prevent bias towards shorter sentences (Papineni et al., 2002):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (5)$$

where p_n is n-gram precision, w_n are weights, and BP is:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \quad (6)$$

- **Cosine Similarity**: Measures the cosine of the angle between text embeddings to assess semantic similarity (Huang et al., 2008):

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \quad (7)$$

where $\vec{A}$ and $\vec{B}$ are vector embeddings of the texts.

4 EXPERIMENTS AND RESULTS

The experiments are detailed in this section, and the results are presented.

4.1 Setup

The system setup features an Intel® Xeon® Silver 4208 CPU, operating at a base clock speed of 2.10GHz, providing robust multi-threading capabilities ideal for handling computationally intensive tasks. Complementing the processor is 62GB of RAM, ensuring smooth multitasking and efficient handling of large datasets. For graphical and parallel computations, the setup includes an NVIDIA RTX 5000 GPU with 16GB of dedicated RAM, which is well-suited for machine learning, deep learning, and high-performance computing applications. The system runs on Ubuntu 20.04.6 LTS, a stable and reliable Linux distribution that offers a secure and versatile environment for development and deployment. The Hugging Face API was employed to access the selected language models (Falcon-7B and LLaMA 2-7B).

4.2 Validation Dataset

Two datasets were used: one for document classification (Dataset 1) and another to evaluate the entire system model (Dataset 2). For training and testing the document classification model (Dataset 1), we used a subset of The RVL-CDIP Dataset (Harley et al., 2015). Although RVL-CDIP originally includes multiple categories, only the "Handwritten" and "Printed" classes were selected. Downsampling was applied to balance both classes, and the data was split into training (75%), validation (15%), and test (10%) sets, each containing the same number of images per class: 4650 images per class for training, 930 for validation, and 621 for testing. For the system model evaluation (Dataset 2), the dataset contains 112 documents of different types (e.g., payment slips, receipts, coupons, etc.). Of these 112 documents, 48 are images of handwritten documents, and 65 are printed. Among the 65 printed documents, 47 are of good quality, while 18 are of poor quality. This dataset can't be made available because it is confidential.

4.3 Experiments

The parameters utilized in the experiments are summarized in Table 2. The table outlines the essential components and configurations of the system, detailing the models, hardware, preprocessing steps, and inference criteria used throughout the process. The experiments were conducted using two state-of-the-art LLMs: LLaMA2-7B and Falcon-7B. Inference was performed on an NVIDIA RTX 5000 GPU with 16GB of dedicated VRAM to ensure efficient processing. Preprocessing involved alignment correction with a threshold of $\pm 5^\circ$, improving the accuracy of text detection. The OCR stage employed the robust Tesseract OCR engine for text extraction from document images. The system was designed to process textual data effectively, with a maximum prompt length of 500 tokens for the LLMs. An ONNX-based image classification model was used to distinguish between handwritten and printed documents, with a confidence threshold of 0.7 to ensure reliability. Additionally, a quality filter was applied during the OCR process, requiring a minimum confidence score of 40 for extracted text to proceed to the following stages. To handle potential inference issues, a maximum of three retries was permitted for each document. The final extracted information was presented in a standardized JSON structure, providing consistent and interpretable results across all experiments.

4.4 Results

To evaluate the proposed system's effectiveness, a series of experiments were conducted using two distinct datasets: one for document classification and another for system-level evaluation. These experiments assessed the model's performance in classifying document types, extracting critical information,

Table 2: Parameters Used in the Experiments.

Parameter	Value / Description
Models	LLaMA2-7B, Falcon-7B
Device	GPU (NVIDIA RTX 5000, 16GB VRAM)
Preprocessing Steps	Alignment correction, threshold $\pm 5^{\circ}$
OCR Engine	Tesseract OCR
Prompt Length	500 tokens
Image Classification	ONNX model, confidence threshold = 0.7
Quality Threshold	Minimum OCR confidence = 40
Inference Attempts	Maximum retries = 3
Output Format	Standardized JSON structure

and identifying anomalies in financial and administrative records. The metrics used for evaluation, including precision, recall, F1-score, BLEU, and cosine similarity, were chosen to comprehensively understand the system's accuracy, reliability, and contextual understanding.

4.4.1 Results Document Classification

Since our dataset contains various types of documents (Dataset 2) and our proposed method is only applied to printed documents, we trained a classification model to recognize whether a document is printed or handwritten (Dataset 1). Table 3 presents the precision, recall, and F1-score.

Table 3: Evaluation Metrics for Document Classification (Dataset 1).

Label	Precision (%)	Recall (%)	F1-Score (%)
Printed (0)	95.84	96.46	96.15
Handwriting (1)	96.43	95.81	96.12

In Table 4, we presented the confusion matrix of document classification. The model can correctly recognize 595 handwriting documents over 26 misunderstood as printed documents in this specific dataset.

Table 4: Confusion Matrix for Document Classification (Dataset 1).

		Predicted Class	
		Printed (0)	Handwriting (1)
Actual Class	Printed (0)	599	22
	Handwriting (1)	26	595

4.4.2 Results LLM

In our experiments, LLaMA2-7B achieved a BLEU score of 0.673 and a cosine similarity of 0.707, demonstrating its ability to generate outputs closely aligned with reference data in semantic relevance and linguistic accuracy. Falcon-7B performed slightly better, with a BLEU score of 0.691 and a cosine similarity of 0.734, highlighting its robustness in producing text that is both syntactically precise and semantically meaningful, making it suitable for high-quality document understanding and summarization tasks.

To contextualize these results, we compare them to values from the literature, such as Yuan and Färber (2023) , where BLEU scores ranged from 0.505 to 0.802, and de Vos et al. (2022) , where cosine similarity scores reached 0.738 and 0.703 in the 'Vehicles' dataset. However, direct comparisons should be made cautiously due to differences in dataset characteristics, evaluation paradigms, and task objectives. While BLEU is traditionally used for machine translation, our evaluation involves different linguistic and contextual challenges, making absolute numerical comparisons less straightforward.

5 CONCLUSIONS AND FUTURE WORK

This paper introduced a system leveraging LLMs, specifically LLaMA2-7B and Falcon-7B, to enhance audit processes in notary offices by automating the extraction and analysis of financial data from various document types. The proposed approach improved transparency, accuracy, and efficiency in auditing by addressing inefficiencies, high costs, and the complexity of unstructured data. The system delivered strong performance in BLEU and cosine similarity metrics, demonstrating its effectiveness in information extraction and anomaly detection. Key benefits include assisting court analysts in identifying fraud cases, optimizing public resource management by eliminating unjustified expenses, and potentially increasing court revenues to reinvest in public services, further reinforcing the system's impact on financial oversight and accountability.

Future work aims to expand the capabilities of the system, particularly in the processing of handwritten documents through handwriting recognition or specialized training. Integrating multimodal learning to analyze text alongside visual elements like stamps and signatures could further enhance its robustness. Additionally, developing multilingual and cross-jurisdictional models would improve the system's adaptability to different languages and regulatory environments, ensuring broader usability and compliance with international standards. Furthermore, domain-specific fine-tuning for legal and financial contexts, real-time auditing features, and improved interpretability would make the system more precise and user-friendly. These advancements will contribute to greater accountability, resource management, and public trust in legal and financial oversight.

ACKNOWLEDGMENTS

The results presented in this paper have been developed as part of a project at SiDi, financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/9.

REFERENCES

Alhamad, H. A., Shehab, M., Shambour, M. K. Y., Abu-Hashem, M. A., Abuthawabeh, A., Al-Aqrabi, H., Daoud, M. S., and Shannaq, F. B. (2024). Handwritten recognition techniques: A comprehensive review. *Symmetry*, 16(6):681.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374.

Beltran, A. (2023). Fiscal data in text: Information extraction from audit reports using natural language processing. *Data & Policy*, 5:e7.

de Elias, E. M., Tasinaffo, P. M., and Hirata, R. (2019). Alignment, scale and skew correction for optical mark recognition documents based. In *2019 XV Workshop de Visão Computacional (WVC)*, pages 26–31. IEEE.

de Vos, I. M. A., Boogerd, G. L., Fennema, M. D., and Correia, A. D. (2022). Comparing in context: Improving cosine similarity measures with a metric tensor. *arXiv preprint arXiv:2203.14996*.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fisher, I. E., Garnsey, M. R., and Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214.

Grabb, D. (2023). The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6.

Harley, A. W., Ufkes, A., and Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. *CoRR*, abs/1502.07058.

Hegghammer, T. (2022). Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5(1):861–882.

Huang, A. et al. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

HuggingFace (2024). Falcon-7b model card. https://huggingface.co/tiiuae/falcon-7b. Accessed: 2024-11-08.

Ingle, R. R., Fujii, Y., Deselaers, T., Baccash, J., and Popat, A. C. (2019). A scalable handwritten text recognition system. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 17–24. IEEE.

Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., and Vassilakopoulos, M. (2023). Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290.

Kumar, P. (2024). Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260.

MetaAI (2024). Llama 2-7b model card. https://huggingface.co/meta-llama/Llama-2-7b. Accessed: 2024-11-08.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rudžionis, V., Lopata, A., Gudas, S., Butleris, R., Veitaitė, I., Dilijonas, D., Grišius, E., Zwitserloot, M., and Rudzioniene, K. (2022). Identifying irregular financial operations using accountant comments and natural language processing techniques. *Applied sciences*, 12(17):8558.

Saini, R. (2015). Document image binarization techniques, developments and related issues: a review. *International Journal of Computer Applications*, 116(7):0975–8887.

Santana, A. F. B., de Faria, J. A., and Sena, T. R. (2024). Editorial volume 05, número 02, 2024.: Auditoria e seus desafios (ainda) atuais! *Revista Controladoria e Gestão*, 5(2):1–3.

Simunic, D. A. (1980). The pricing of audit services: Theory and evidence. *Journal of accounting research*, pages 161–190.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., and Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7(1):41.

Yuan, S. and Färber, M. (2023). Evaluating generative models for graph-to-text generation. *arXiv preprint arXiv:2307.14712*.

Zheng, X., Zhang, C., and Woodland, P. C. (2021). Adapting gpt, gpt-2 and bert language models for speech recognition. In *2021 IEEE Automatic speech recognition and understanding workshop (ASRU)*, pages 162–168. IEEE.