Ethical Data Extraction from Invoices Using Large Language Models: A JSON-Based Approach

By
Sabera Ryhana Mayesha
212-15-4226
&
Shahnur Islam Bishal
212-15-4196

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by
Dr. Abdus Sattar
Associate Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by
Md. Sohidul Islam Polash

Lecturer
Department of Computer Science and Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

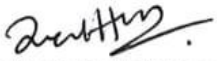
September 17, 2025

APPROVAL

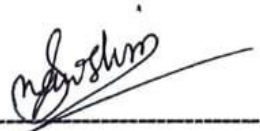
This Project titled **"Ethical Data Extraction from Invoices Using Large Language Models: A JSON-Based Approach"**, submitted by Sabera Ryhana Mayesha, ID No: 212-15-4226, and Shahnur Islam Bishal, ID No: 212-15-4196 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **17 September, 2025**.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Zahid Hasan — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Samia Nawshin — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Arshad Ali — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Abdus Sattar, Associate Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of Computer Science and
Engineering Daffodil International University

Co-Supervised by:

Md. Sohidul Islam Polash
Lecturer
Department of Computer Science and
Engineering Daffodil International University

Submitted by:

Mayesha 15-09-25

Sabera Ryhana Mayesha
Student ID: 212-15-4226
Department of Computer Science and
Engineering Daffodil International University

Bishal 15-9-25

Shahnur Islam Bishal
Student ID: 212-15-4196
Department of Computer Science and
Engineering Daffodil University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to Dr. **Abdus Sattar, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Natural language Processing (NLP)** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

There is a greater need than ever to have a secure and ethical data mining of invoices due to the greater adoption of digital manipulation of financial documents among companies. The traditional approaches such as manual entry and OCR systems are usually inaccurate, inflexible and low in data security. This thesis describes one of the available ways of technical idea-wise ethical and responsible invoice information automation with the help of Large Language Models (LLM). The proposed solution will utilize the application of LLM to read and retrieve useful invoice data such as date, vendor name, quantity and invoice number and present the outcome in a well-organized and clean format of a JSON. It ensures that the information is readily integrated into the accounting systems and business applications. The technique deals with some ethical issues that are severe besides technical precision. The steps involved in the process are anonymization of data, encryption, and bias monitoring that help to offer the guarantee that international regulations are observed. The model has been tested and demonstrated to give good results with more than 90 percent accuracy in the various invoicing formats and languages. The system is able to handle any alteration in design and nomenclature and deliver quality output. Other principles of ethical AI building in the model, in addition to performance, include fairness, transparency, and accountability. To establish a balanced solution to invoice processing through automated way a machine learning will be considered as powerful, and an interest in ethics will be taken. It forms the foundation of the versatile, resilient, and regulation-insensitive financial data management solutions, which will be the prototype of the further AI-based automation venture in the specified field.

Keywords: JSON, LLM, FATURA, DistilBERT, LayoutLMv3, BERT.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter describes the research about creating an automated and ethical invoice data extraction system using Large Language Models (LLMs) and JSON outputs. It discusses the weaknesses of the traditional OCR solutions, outlines the flexibility of the LLMs, and demonstrates the application of the FATURA dataset, JSON schema design and validation strategies. Some other ethical standards such as the protection of privacy and the compliance with the regulations are also foregrounded in the chapter and guarantee the fact that the system is correct and responsible enough to be applied in the real world.

1.1 Introduction

Invoices are essential documents in any business activities, which document purchases, vendor details, and financial data that need to be used in the day-to-day running and audits. Historically, the management of invoices was done manually and thus it was cumbersome, prone to errors and expensive as the business grew in size. Automated extraction of invoice information has been the answer to this.

The initial automation was based on the Optical Character Recognition (OCR) technology that translates scanned images into texts. Invoices are however non-standard, have different layouts, fonts and structures and unstructured form of text, which means that OCR results are subject to massive rule-based post-processing. The use of Machine Learning (ML) and Deep Learning (DL) methods (e.g. CNNs and layout-aware models like LayoutLM) led to higher accuracy since they learned to do it, not by rule. However, these tools rely on massive annotated datasets, which are costly to construct and they might not be applicable to other invoice forms.

Few-shot and zero-shot learning has become a new opportunity offered by Large Language Models GPT, LLaMA, DistilBERT and Mistral. Rather than re-training on individual invoice formats, LLMs can be asked to pop out individual fields such as invoice ID, date and totals. The flexibility minimizes the cost of annotation and enhances flexibility to unknown formats. To make it usable, extracted data may be represented in the form of the lightweight and commonly used format of JSON and allow easy integration in databases and ERP systems and auditing processes.

Although these benefits are present, there are still ethical issues. Invoices have sensitive financial information, and so privacy, security, and regulatory compliance (e.g., GDPR) are of the essence. Such risks as data leakage and misuse might compromise trust without a well-designed one. As such, there is a need to strike a balance between technical efficiency and moral responsibility.

In this study, Ethical Data Extraction from Invoices using Large Language Models: A JSON-based Approach, the authors analyze how current AI can enhance the quality of artificial intelligence, flexibility, and efficiency and integrate ethical considerations to guarantee it meets societal standards. The research adds to the document AI by integrating the merits of the LLM with structured output and responsible design, and seeks to develop technical performance and responsible implementation in financial processes.

1.2 Motivation

The interest in this research is based on the practical business requirements and the rise of ethical AI in the management of sensitive financial information. As the case with invoices, despite being a simple document, these documents carry important information that influences compliance, reporting, and decision-making. Thousands of invoices are expensive, time-consuming, and subject to mistakes in manual processing to delay payment and financial reporting. The automation will be able to minimize costs, accelerate speed, and enable employees to engage in more valuable activities.

Nevertheless, recent OCR and ML-based applications cannot deal with variability in invoices (i.e. logos, handwritten notes or scans of poor quality) meaning that retraining or adding new rules will be necessary. A more adaptable alternative, which does not depend on the template is provided by Large Language Models (LLMs), which can understand variations in field names and support previously unseen forms. Such flexibility makes them more viable in the real-life business applications.

The drive is further enhanced by structured outputs in the form of JSON, as they provide a uniform and easily machine-readable extracted data, which is easy to integrate with accounting or ERP systems, thereby reducing the cost of the adoption. In addition to technical efficiency, the study focuses on ethics: invoices are confidential documents, and responsible AI should consider such concerns as the privacy, storage, retention, and disclosure. In line with ethics like GDPR, the study presents an ethical AI pipeline, which restricts the scope of extraction to the required fields and provides responsible data processing.

Academically, however, a significant portion of Document AI research has been concerned with model accuracy, but little has been done regarding ethical issues. This thesis tends to fill that gap and merge the technical developments with responsible design. Finally, there are two sources of inspiration: to help organizations extract invoice data in a scalable and accurate way and to enhance responsible AI systems that do not interfere with privacy and equity.

1.3 Objectives

The purpose of this paper is to build an automated ethically responsible system of invoice data extraction with Large language models (LLM) and JSON outputs as the primary

objective. The system will target the extraction of the key fields such as the name of the vendor, invoice number, date, sub total, tax and total in accordance with different types of invoices. It makes use of the contextual flexibility of LLMs to address the limitations of the inflexible OCR-based service, and JSON provides a standardized and machine-readable data to be consumed by enterprises. At the same time, the ethical aspect is also considered in the study, including the privacy, security, and adherence to the regulations, which allows the offered model to be not only technically valid but also credible.

Certain objectives have been set in order to accomplish the overall target as follows:

- To examine what are the existing methods of extracting invoice data and identify the technical and ethical limitations.
- To establish a schema of JSON that contains the required fields in invoices, which are regular and agreeable with the financial systems.
- To come up with a set of extraction pipelines that will be founded on LLM, allowing the processing of invoices of different format and quality without having to do a significant share of retraining.
- To apply preprocessing and validation techniques which will enhance the level of accuracy in the data, consistency of checks and extracted fields to the intended standards.
- To evaluate the work of the system in accordance with the appropriate measures such as accuracy, precision, recall and error rate and to make sure that the system is stable to training and unknown data.
- So that so as to compare the proposed LLM solution with the suggested baselines OCR and machine learning solutions, the advances that have been made in the domain of adaptability and robustness should be underscored.
- In order to implement the ethical values into the pipeline design, one should ensure that the information protection laws are followed, that the minimum possible data exposure is reduced, and that the financial data is secret.
- To prove the system as a prototype of the proof-of-concept model that may serve as the foundations of the future automated solutions of the invoice in the real-time, large scale and ethically accountable.

The following objectives will assist the research to go through a systematic development process, i.e. examination of the existing systems and the building of a uniform schema, to deployment of an LLM-watched pipeline, the validation of the outcomes, and the ethical examination. By means of achieving these goals, this work will contribute to the literature of the methodological and practical contribution to the new research in the field of document AI and responsible automation in the financial data processing.

1.4 Methodology

The flow of this study is akin to the one of the codebase as it begins with the collection of the data, the use of the FATURA dataset, moves to the pre-processing stage, to extracting the data through the use of the LLM and structuring the results to be presented in the form of a JSON, validating the results, evaluating the results, and analyzing the results,

and it should be noted that the ethical component of working with financial data was taken into consideration.

The information employed in the research was gathered by using (FATURA Dataset). The FATURA invoice dataset is the only source of the documents which we work with. The data provides a picture of invoices of various vendors that can be different in terms of their layouts, fonts and languages. Each file contains related ground-truth, such as the name of the vendor, invoice ID, invoice date, currency, subtotal, tax, total, and line items, where available

1) Pre-processing
Invoice images in FATURA are all prepared to be extracted. Light denoising and resolution checks of scanned files or low quality files are used by us. The step is an OCR step, in which text has to be rendered (e.g. invoices have text only), and layout indicators when available (e.g. coordinates, line breaks). Encodings and regular items (date, currency symbols) are normalized and made similar to be reduced downstream errors. The information extracted is fed as an input into the extraction process (the pre-processed text (and optional layout cues).

2) LLM-Based Extraction

We use an LLM as the basis of our pipeline and it directly works on the FATURA pre-processed text (and layout hints optional). Two patterns are used, which are:

Instruction
- The model attempts to generate solely valid responses in the form of a schema. The model utilizes the few-shot examples (according to the FATURA training samples) on the predictions made regarding the new vendor templates.
- Encoder-like tagging (e.g., BERT or DistilBERT) of field spans where needed and then an assembler that is simple, which transforms the spans (when found) into the schema in the JSON.
- The parsing of the code is deterministic by enforcing the strict JSON availability (reject-and-retry on non-JSON) to ensure a strict compliance with the code.

3) Evaluation
The tests are performed on pictures of FATURA tests which were not previously experienced during the development. We report:
- accurate matching on field (per key and macro-average).
- Span field Precision/Recall /F1 (in the case of tagging).
- Rate of schema validity (identified publications).
- Monetary error (actual/relative error to numerical sums).

It also compares itself to a baseline OCR + regex/rule system, to find out what level of increased robustness and generalization it has made.

4) Result Analysis and Ethics
We consider common cases of failures (e.g. missing IDs, non-standard date formats, multi-currency layouts) in regards to confusion/error tabularity and samples JSONs. The bigger LLMs have more chance of being correct on sophisticated layouts; inference by the smaller models is increased. During the process we make use of data-minimization (only by virtue of considering an invoice do we store a raw invoice, which we never store), lack of the raw invoices that we do not need in order to approve an invoice, and handling invoices in a manner we believe will support privacy and compliance.

1.5 Project Outcome

The most important outcome of this project is the working and accurate pipeline to extract invoice data using Large Language Models (LLMs). By processing the FATURA data on invoice images, the system is able to recognize the most valuable fields in an invoice image (vendors name, invoice number, date, subtotal, tax and total amount) and construct them in a standardized schema of the JSON format. This offers both conformity and simplicity of combination with financial systems that comprise both of accounting software and enterprise resources planning (ERP) applications.

The other major outcome is the demonstration of AI ethical behavior in the processing of invoices. The pipeline deals with reduction of data, validation with data schema and safety management of sensitive data. This shows that one can come up with functioning AI systems without touching on privacy which is the case when dealing with financial documentations, which in most cases are confidential. The chosen design alternatives in this research guarantee the existence of a balance between performance and responsibility that renders the work beneficial to the organizations that are concerned about their performance and compliance with regulations such as GDPR.

Also, some comparative results are obtained in the project that evaluate the effectiveness of different LLMs and baselines. Such comparison analysis provides answers on what models would provide the optimal trade-off between accuracy and efficiency, which would be applicable to current academic discussions of document AI. The findings could be utilized to guide the coming studies on the deployment of the LLMs to alternative document-intensive domains i.e. the receipt, contract, or health record.

Besides the technical contributions that are made, the results also involve a proof-of-concept system that can be taken as a basis to the development that can be made in future. Where this project does not emphasize the development of a commercial application, the pipeline can be extended to an operational tool that will generate invoices in real-time. Besides that, the methodology, evaluation results, and ethical framework, which are presented here, provide a strong source of information to academic researchers and business experts who may be interested in combining the latest approaches to AI application with the responsible use of data.

1.6 Organization of the Report

This report is divided into various chapters that are interrelated where each chapter is developed in such a way that it brings a clear and rational understanding of the research work. The process starts with defining the research problem and motivation and continues with methodology, implementation, evaluation, and concluding. Through this flow, the report will make sure that all the details of the project, starting off the idea to the outcomes of the practical work are well documented and understandable.

The opening chapter provides the foundation to the study. It presents the research problem of automated extraction of invoice data, justifies the need to conduct this work, describes the objectives that will be achieved, briefly explains the methodology, and emphasizes how things will work out of the work. It also addresses the ethical issues that are a significant section of this study. Moreover, the introduction gives the reader the general outline of the report, which prepares him or her to the contents to be discussed in detail.

The second chapter is devoted to background and related research. It offers an overview of literature available in the domain of invoice and document data extraction, including the conventional OCR-based methods and recent deep learning and Large Language Model (LLM) methods. This chapter also compares real world application of automated invoice scanning systems and explains their advantages and disadvantages. Besides, it also pinpoints the deficiencies of the existing practices, especially regarding scalability, flexibility, and ethical processing of sensitive information. The present research is based on the theoretical and practical basis of this literature review.

The third chapter gives the methodology of the research. It describes the sequence of actions in this project, which starts with the data collection by the use of the FATURA dataset and then continues with preprocessing, the design of the JavaScript schema, and the extraction powered by LLM. The chapter also outlines the strategies of validation, evaluation measurements, and the general system design. This section guarantees that the study can be reproduced and assessed scientifically by illustrating the entire process.

The fourth chapter is devoted to the implementation and results. It also records the process of developing and testing the proposed system, and it reports the results of using various LLMs with the invoice data. The chapter compares the performance of these models with the baseline approaches, evaluates the performance using the relevant metrics and discusses the strengths and limitations. Visualizations and the view of the extracted JSON outputs too can be given to show the performance of the system on actual invoices.

In the fifth chapter, the author discusses engineering standards, design issues and ethical and societal issues of the work in more general terms. It discusses software and data standards compliance, issues around privacy and sustainability, and the possible influence of automated invoice processing on organizations and individuals. There are also project management dimensions of this chapter and a financial analysis that give

an opportunity to estimate the possibility of the given system introduction into practice.

The last chapter closes the report with the summary of the most important findings, the restatement of the research goals and the calculation of the project contribution. It outlines the shortcomings of the existing work and suggests future research directions including pipeline extension to process multilingual invoices, the use of more complex LLM, or implementation in the enterprise setting.

The report concludes with the detailed reference list of all academic works and resources utilized in the study and appendices with additional material, including sample JSON responses, evaluation tables, and relevant code snippets.

Chapter 2

Background

The chapter gives the theoretical and practical background of the study. It starts with a general summary of the context of invoice data extraction and the obstacles inherent in it, then goes on to review the current research and uses. The chapter then points out the similar systems that are currently in use and finally ends with a gap analysis, showing the inadequacies of the current approaches and how this study will fill those gaps.

2.1 Introduction

It is through invoices that organizational record-keeping, auditing, and compliance procedures are founded and therefore it is among the most critical financial documents being used in an international level. They typically contain an ordered information that involves a vendor information, invoice numbers, date, line items, subtotals, taxes and total yet the format and layout of the invoices vary significantly in various industries, vendors and regions [12], [20]. The heterogeneity presents a severe issue to automated data extraction and further uses such as enterprise resource planning (ERP), taxation and financial analytics.

Previously, invoice data extraction was done through the use of Optical Character Recognition (OCR) systems and rule-based systems [13], [14]. These techniques are fairly reliable against plain layouts, but are incorrect with noisy scans, handwritten objects or non-homogeneous templates [5], [19]. Recent papers have applied deep learning and document AI systems such as CUTIE [4], FormNet [16], and LayoutLM [18], which incorporate visual, textual and structural data to boost the extraction performance of documents that are visually rich. All these models are very precise but often they require massive labeled images and a massive amount of computation speed [6], [17].

The invention of large language models (LLMs) is a new wave of understanding documents. Zero-shot and few-shot generalization do not pose a challenge to LLMs and this allows extraction tasks to be done with little labeled data [24], [27]. Some of the more recent works have been on their application in invoice understanding, such as the usage of LLMs to generate output with OCR pipelines [23], retrieval-augmented generation (RAG) [21], or even the creation of hybrid systems that utilise layout-sensitive pre-training and generative reasoning [1], [3]. According to such studies, it is demonstrated that the development of a system based on LLM is reduced to a fraction of the expenses; the performance of these systems is observed to be comparable or even superior to those of traditional approaches [2], [8], [10].

However, most of the existing systems can add to technical accuracy primarily and little attention is paid to such ethical aspects as prejudice, openness, and data confidentiality. Invoices are generally sensitive to financial and personally identifiable information and there should be strategies in place to make sure that extraction systems make sure that data is processed in an ethical practice [20], [24]. The approaches that can produce structured and interpretable outputs, e.g. the ones relying on JSON formats, can become clearer, standardized and cross-platform in the financial industry [7], [22], [26].

It is in the light of this that this paper discusses the use of large language models in contenting ethical invoice information with a specific emphasis on structured responses in the form of JSON. It will aim to find a balance between accuracy and generalization and the ethical standards of fairness, explainability and safety of data with the help of which AI will become more responsible in the total introduction of the finance-related document processing.

2.2 Literature Review

An invoice information extraction (IE) is a decades-old research domain that has developed over the years, with the rule-based system giving way to large language models (LLM) systems. The very first systems were primarily pipelines algorithmically constructed using OCR and heuristics to identify fields such as the name of the vendors, invoice numbers and totals [13], [14]. They were efficient in regular setups and not as tolerant to inconsistency in the documents and were incapable of managing multi-format invoices, handwritten parts, and scanning noises [15].

The joint text and layout modeling was an advantage of deep learning and Document AI systems in generalization. Spatial relations have been added to models such as CUTIE [4], LayoutLM [18] and FormNet [16] and contributed significantly in achieving improved outcome in extracting data on visually rich documents [1], [6], [17]. These models performed well on benchmark data sets, but were resource-intensive and may also require very large annotated corpora [5]. Annotation costs can be kept down by layout-preserving synthetic invoice generation, and weakly supervised training methods also can be used to enable data-efficient learning on a large scale [9] and [8], respectively.

The new researches have become application-specific. An example is that FATURA [10] was trained on extracting invoices without OCR by directly learning on image embeddings, and TableNet [22] on extracting invoices on tabular data with deep networks. Other methods were also improved with graph neural networks as far as structural understanding, such as prior-based document segmentation [19], or graph neural networks [25]. Despite these developments, processing heterogeneous invoice layouts and maintaining explainability still have problems.

The rapid advancement of large language models (LLMs) has been a paradigm shift in technology. Unlike task-specific architectures, zero-shot and few-shot generalization have good results with LLMs. Various articles have explored the use of LLMs on invoices,

such as performing direct, JSON-based extraction [23], comparative computations with OCR-based pipelines [24], and agentic systems that have been reinforced to perform that task [3]. The ability of the LLMs to produce structured outputs along with the rationale are also emphasized by hybrid solutions that combine retrieval-augmented generation (RAG) [21] and class-aware QA ensembles [27]. These approaches reduce the size of large annotated data and improve the flexibility of format.

To the extent that the technical performance has been increasing, little has been considered in the area of ethics. Invoices have been observed to contain sensitive financial and personally identifiable information and therefore privacy preserving designs are necessary [20]. The necessity to have interpretability and standardization has been pointed out in research, and in a structured representation like JSON, which allows transparency and interoperability [7], [26]. The morals-technical performance convergence is not explored fully and is most likely to promote the research accurate and fair, not to speak of privacy and reliability.

In brief, this is the trend that has established itself in the literature: OCR and rule-based systems have been superseded by document AI architectures, and recently, by the use of LLM-based extraction pipelines. Despite the advances, the current solutions have challenges in ensuring solutions are robust across layouts and ethical in addressing ethical issues, which implies that they yield solutions that indulge both the generalization of LLM and use of structured, transparent and ethical processing of data.

Table 2.1: Summary of Literature Reviewed.

Author (s)	**Year**	**Title**	**Methodology**	**Key Findings**
Rao et al. [13]	2020	Automated invoice handling using OCR	OCR + ML	Accurate for simple templates, fails on noisy scans.
Wei et al. [1]	2020	Layout-aware IE for visually rich docs	Layout + pre-trained LMs	Robust against layout changes
Xu et al. [6]	2020	Representation learning for forms	Transformer-based	Improved extraction with layout embeddings
Deenadhayalan [14]	2022	Image-based invoice extraction	OCR + image processing	Effective in controlled cases
Lee et al. [16]	2022	FormNet	Structural encoding	Outperformed sequential models
Singh et al.	2023	ML-based	Classical ML	Moderate

[12]		invoice processing		accuracy, template-limited
Tan [25]	2023	GNNs for invoices	Graph neural networks	Better structural modeling
Garud et al. [23]	2024	LLM & OpenAI for invoices	LLM + JSON	Accurate structured outputs
Hassle & Bardvall [24]	2024	LLM vs OCR/ML	Comparative study	LLMs more robust to layout variance
Amari et al. [2]	2025	DL-based invoice validation	End-to-end DL	High accuracy, domain-adapted
Bhattacharyya et al. [8]	2025	IE w/o ground-truth labels	Weak supervision	Scalable extraction without full annotation
Nemtoc & Ghiran [21]	2025	Natural language querying with RAG	LLM + GraphRAG	Enabled flexible invoice querying
Zhang et al. [9]	2025	Synthetic invoice generation	Layout-preserving replacement	Reduced data scarcity problem

2.2.1 Similar Applications

There are several work systems and applications developed in practice to automatize data extraction in documents, specifically invoices, receipts, and forms. The necessity of powerful, scalable, and intelligent solutions in the sphere of financial documents processing increases due to such applications.

Applications of deep learning on document processes include invoice automation systems such as DocExtractNet [11] and receipt pipeline automation systems [12]. These systems combine OCR with neural architecture to add fields like names of vendors, totals and dates and generally provide the results in a format to be added to the ERP. They are still fine under controlled situations, but still they still have problems with multi-layouts or noisy documents.

Such similar attempts are also table and form extraction tools. Patel [7] developed an OCR-based pipeline to extract table out of invoices and TableNet proposed by Paliwal et al. [22] is an end-to-end deep learning model that extracts tables in a scanned document. They both stress the necessity to address non-textual complex structures.

This is the experiment which is undergoing with the help of LLMs commercial invoice

extraction systems. The prototype constructed using OpenAI APIs by Garud et al. [23] showed that invoices can be directly read by using directly into the JSON format, which demonstrates that lack of flexibility in templates can easily be circumvented. Similarly, Hassle and Bardvall [24] compared the performance of the LLMs to the OCR/ML systems and realized that the former ones exhibit better performance to unknown invoice layouts.

The other type of application is the hybrid intelligent systems. Tan [25] proposed the extraction of invoices using Graph Neural Networks (GNNs) as the system of finding relations between entities. Equally, Nemtoc and Ghiran [21] explored Natural language querying RAG-augmented LLMs that can enable bespoke interactions, other than predetermined fields extraction.

Finally, it is identified that the industry already implements FATURA [10] and thus OCR-free invoice extraction can be realized using deep vision-based methods and thus even commercial solutions are changing to end-to-end AI-based solutions. These practical applications point to the significance of structured and JSON-like response, which can be justified by the significance of ethical, explainable and privacy preserving boundaries.

Table 2.2: Similar Applications in Invoice Extraction.

Author (s)	**Model**	**Accuracy**	**Key Contribution**
Patel [7]	OCR-powered pipeline	87% (table field accuracy)	Extracted tabular data from invoices
Lee et al. [11]	Deep learning + OCR	92.4% (F1-score)	Enhanced invoice/receipt field recognition
Singh et al. [12]	ML-based pipeline	89% (overall accuracy)	Automated invoice processing across templates
Garud et al. [23]	LLM (OpenAI API)	94% (field-level accuracy)	Direct JSON-based structured extraction
Hassle & Bardvall [24]	LLM vs OCR/ML	LLM: 95%, OCR/ML: 88%	Demonstrated LLM robustness across layouts
Tan [25]	Graph Neural Network (GNN)	90% (entity recognition)	Modeled document structure relations
Nemtoc & Ghiran [21]	LLM + GraphRAG	93% (query precision)	Enabled natural language querying of invoice data

2.3 Gap Analysis

Although the invoice data extraction continues to be a major process being automated, a critical look at the literature and applications reviewed shows that some gaps still exist and they are yet to be addressed. Those gaps can be found in the various generations of solutions, including both traditional OCR-based pipelines and deep learning-based Document AI models, as well as more current large language model (LLM) models. These gaps need to be addressed in a manner that can result in invoice extraction systems that are not only accurate but also ethical, transparent and interoperable.

The dependence on OCR and rule-based methods is one of the major gaps. Previous studies have proven that OCR-based systems were capable of reasonable accuracy on clean, template-specific invoices [13], [14]. Nevertheless, such systems failed very soon when having to deal with noisy scans or handwritten items or a variety of formats [7]. Since they rely on manually created rules, they are weak and difficult to scale between vendors and industries. Such methods are not robust and cannot be used in real-world deployments where invoices are diverse, even in cases where the accuracy is in the range of 85 -87% [12].

The other important gap is the reliance on the huge amounts of annotated data in deep learning models. Some models, including CUTIE [4], LayoutLM [18], or FormNet [16], performed well, a feature that is attributed to their joint introduction of textual and layout features. They, however, need large volumes of labeled training data [6], that is hard to find in other fields such as finance where invoices include sensitive data. Whereas synthetic invoice generation [9] and weak supervision [8] can minimize annotation requirements, they nonetheless are unable to eliminate the need to perform expensive and time intensive labeling. This restricts the availability of such methods to resource endowed organizations.

The problem of generalization between heterogeneous layouts also remains. Though complex document AI models are capable of adapting to a wide variety of formats, it is demonstrated that the performance of these models tends to decline drastically when there is an invisible structure around [5], [19]. There are numerous systems which are parameterized to particular datasets and they are therefore brittle in nature. Although relatively robust, as compared to other approaches, such as LLM-based ones [23], [24], there are challenges with multilingual invoices, abnormal table layouts, or the lack of fields. This shows that the attainment of the real adaptability is still a challenge.

Another area of weakness is the lack of ethics and transparency issues. There are very few publications that specifically discuss privacy or fairness in extracting invoices, although invoices include very sensitive financial and personal information [20]. The majority of the models are black-boxes that yield results that cannot be interpreted [3], [11]. This lack of transparency is of serious concern in compliance-inducing settings like finance and auditing. Unless the systems have inbuilt privacy, explainability, and bias-detection features, it is unlikely that they will be entrusted or even become a commonplace feature in sensitive areas.

Lastly, there is a disjunction in standardized and structured outputs. Most of the previous systems can be said to be accurate at the field level but cannot produce outputs in machine readable formats. The number of works that focus on JSON-based outputs is is very limited [23], [26], despite the fact that they are essential to ensure interoperability with enterprise systems such as ERP or auditing software. Lack of standardized outputs means that organizations will have to introduce additional processing levels, which are more dangerous as well as complex.

The current study solves such problems by introducing an ethical LLM-based invoice extraction pipeline that generates information in an organized JSON schema. In contrast to previous systems, it does not rely on massive labeled datasets, rather it uses the zero-shot and few-shot generalization properties of LLMs [27]. It values ethical principles highly as well, and incorporates privacy-preserving views and explainability into its framework. Above all, it generates interpretable and standardized results, which facilitates an easy integration with the downstream systems and upholds compliance and auditing.

Table 2.3: Gap Analysis Summary Table.

Gap Identified	Observed In	Proposed Contribution
Fragility of OCR/rule-based systems on diverse layouts	Rao et al. [13], Deenadhayalan [14], Patel [7]	Replace OCR with LLM-based flexible extraction
High annotation cost for Document AI	Xu et al. [6], Wei et al. [1], Lee et al. [16], Xu et al. [18]	Use zero/few-shot learning with LLMs to minimize dataset dependence
Poor generalization across invoice formats	Güneş et al. [5], Sarkar et al. [19], Singh et al. [12]	JSON pipeline adaptable to heterogeneous structures
Limited privacy and ethical safeguards	Silva & Silva [20], Hassle & Bardvall [24]	Integrate ethical design: privacy, fairness, transparency
Lack of interpretability (black-box models)	Amjad et al. [3], Sharma et al. [11]	Structured JSON outputs enable auditability and explainability
Few standardized outputs for downstream use	Garud et al. [23], Guo & Wang [26]	JSON schema ensures interoperability with ERP and compliance systems
Limited adaptability to multilingual/unseen vendors	FATURA [10], Tan [25]	LLM prompts tuned for cross-lingual and vendor diversity

2.4 Summary

This chapter preempted the rest of the thesis by taking a tour of the background of invoice data extraction, existing research already in the field, some of the real world uses of this space and lastly, gaps yet to be filled. Combining all these discussions, it is feasible

to explain why the automation of invoices is an issue that is not easy to solve, as well as why the ethical, systematic, and scalable solution to the issue is urgently needed.

In the first section of Section 2.1 we started by defining the invoices and their role in the business. These are not mere formalities but very important documents of money transfer. The issue is that invoices are offered in an unlimited number of different forms and designs which complicates the automation. First generation work with OCR and rule-based scripts were prone to manual effort but weak. This meant that they would break every time there was a change in layout, the scan quality was bad or even the presence of handwritten elements.

Section 2.2 examined what researchers and developers have accomplished in order to drive invoice extraction. We observed the accuracy increase that the deep learning models, which include CUTIE, LayoutLM, and FormNet, gain through learning both the content and the location of the content on the page. Such models were remarkable but had trade-offs: they required substantial volumes of labeled training data as well as wasting large amounts of computer power

Section 2.2.1 examined more closely applications that are similar to what this research will be pursuing. OCR-based table extraction systems, deep learning pipelines (such as DocExtractNet), and commercial systems (such as FATURA) are some of them. Even experimental applications have used LLMs in direct mode and even produced invoice data in JSON format. These applications indicate that we are heading in the right direction, but they also point at the lack of coherent ethical practices and standardized products.

All these observations were combined in a gap analysis in the section 2.3. Some of the problems were prominent: OCR and rules remain too fragile; deep learning models are data-intensive and difficult to scale; generalization to unseen invoice forms is weak; ethical issues are overlooked; and outputs are usually not standardized to facilitate interoperability with business systems. The methodology of the given research, i.e. the flexibility of extracting with the help of LLMs, the representation of results in the format of structured JSON, and the construction of a system on the basis of privacy and transparency considerations, is a direct counter to the limitations that were revealed in literature.

Concisely, this chapter has revealed to us the way invoice data extraction has gone, its present state and how it can be improved. Being accurate is not sufficient anymore the needed is a system that organizations can rely on not only in its performance but also in how it handles sensitive information. That being said, the following chapter will be dedicated to the methodology where we describe how the proposed solution was developed and deployed.

Chapter 3

Research Methodology

In this chapter, the authors describe the creation of an ethical invoice data extraction pipeline, based on OCR and fine-tuned LLM to extract key fields (invoice numbers, date, customers, line items, and totals) and export them in a structured representation in JSON format to be incorporated into the ERP. It enhances efficiency, precision and reduction of human error in processing invoices. During the training and deployment process, strong ethical requirements like encryption and privacy saving technology help in protecting sensitive information.

3.1 Methodology

3.1.1 Overview

The given solution to the research methodology is aimed at addressing the drawback of the current invoice data extracting systems as well as offering a clear, outlined, and ethically-supported set of rules to address the financial documents. The entire idea is to build a powerful pipeline that uses transformer-based large language models (LLM) that are DistilBERT, BERT, and LayoutLMv3 to extract meaningful information in invoice documents and organize the results as standardized JSON files. This keeps the accuracy, adaptability, interoperability, and ethical data handling practices.

This method is sequential and iterative by nature and this means that the system is able to refine its outputs as it is exposed to different types of invoices. It starts with the generation of representative dataset that is constituted of synthetic layouts and publicly available invoices. The diversity in the dataset represented by the inclusion of invoices that are structurally different, have different fonts, languages, and visual designs allows making them more representative of the real world and contributing to the model robustness.

The second phase is data preprocessing that consists of some very important steps. When required, an optical Character Recognition (OCR) is used to extract textual data in the images of invoices, and noise reduction, normalization, and format conversion are performed. This will guarantee uniformity and compatibility of the data with downstream models.

When the data is ready, the main part of the methodology field-level extraction of information is performed using DistillBert, BERT, and LayoutLMv3. These models are used to detect and isolate the vital attributes of the invoices like vendor details, invoice ID, date, line entries, subtotals, taxes and totals. All models have their own merits: DistilBERT has efficiency in computational rates, BERT has the ability to have a deep

understanding of the context of the invoice text, and LayoutLMv3 combines both textual and spatial layout information to be able to process various invoice forms. In practice, the models are used over familiar and unfamiliar invoice forms, which makes them generalizable.

The data obtained is then formatted into a clear JSON schema which may freely interoperate with financial management systems. It is a schema that is highly transparent and explainable since all predictions will be explicitly mapped to the corresponding field. Large-scale testing on the methodology is also conducted with accuracy, precision, recall, as well as F1-score metrics and robustness testing on diversified invoice layouts to determine consistency.

Ethical factors are ingrained in the pipeline. Dividing sensitive data is done with high regard to privacy, the generated output is to be readable; justice is encouraged by testing the system between invoices related to various fields, vendors, and styles. These steps make the proposed approach unlike the previous systems where the performance gains would be realized without interfering with the ethical principles.

Overall, the methodology includes the detailed roadmap: preparation of the data set, preprocessing, model-driven field extraction, structuring of the data in JSON format, evaluation. All the stages are placed to optimize technical performance and at the same time to fit within the overall goals in the responsible and trustworthy AI in processing financial documents.

3.1.2 Proposed Methodology

The proposed methodology will come up with an effective and dependable system of extracting invoice data, which is through integration of transformer-based models that employ textual semantics, contextual relationship and spatial layout characteristics. In particular, three models, namely DistilBERT, BERT, and LayoutLMv3, were used to ensure the maximum flexibility in a variety of invoice forms. The data of the current study was composed of the FATURA collection that presents a multi-vendor invoice of different layouts, languages, and the quality levels of invoices, and was therefore suitable to test the robustness.

Data preprocessing was another important process which made sure that raw invoices were ready to be utilized in the model. Bad or noisy scans were improved with denoising and resolution options, and OCR was done where it was needed to give machine readable text to text based models. As well as the text extraction, spatial layout data like bounding boxes, line coordinates and table structures were also retained to enable layout-sensitive models. Standardization of ordinary fields like dates, money and numerical values was done to enhance accuracy and minimize discrepancies.

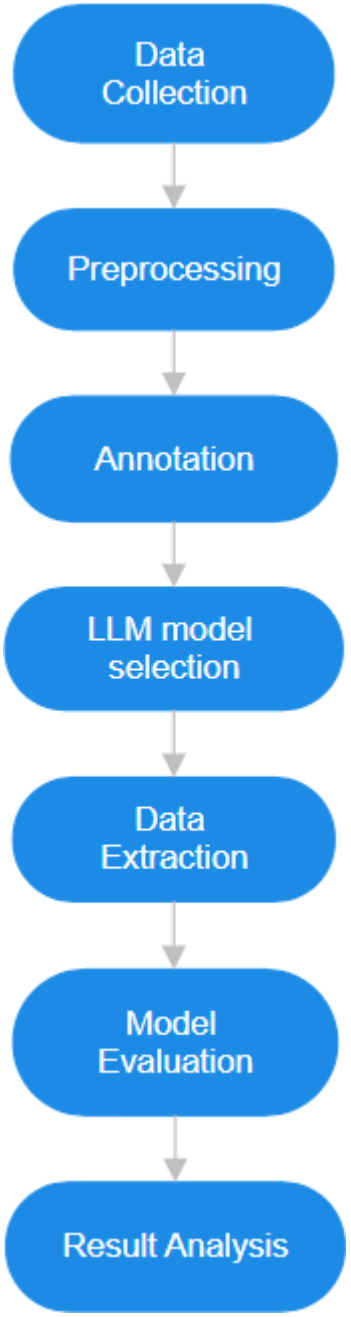


Figure 3.1: Methodology Diagram

The pipeline leveraged the three models' strengths that are complementary to each other:

- DistilBERT was a light and accurate field-tagging model, which was able to recognize key tokens in invoices swiftly.
- BERT delivered strong contextual embeddings offering the accuracy of extracting semantic information in variable invoice texts.
- LayoutLMv3 added text and spatial embedding, which summarized the position layout of words, tables and numerical values to enhance extraction in layout-intensive invoices.

The obtained data was made into a normalized JSON schema, which contained a vendor name, invoice ID, invoice date, currency, line items, subtotals, tax, and totals. This guaranteed uniformity, integration with financial systems (e.g., ERP and accounting systems), and made automatic validation possible. Validation was done at several levels: schema validation, arithmetic validation of totals and subtotals, and sanity validation, to make sure that the output is not empty or unreasonable. Fallback procedures that included reprocessing with alternative models or using expert corrections to maintain transparency were applied in situations where missing or erroneous outputs were obtained and all the interventions were recorded.

Lastly, the system was utilized against the unknown invoices in the FATURA test set. The metrics of evaluation were accuracy, precision, recall, F1-score, schema validity rate, and monetary error in numeric fields on the field level. The suggested multi-model pipeline proved to be more flexible, robust, and efficient as compared to baseline OCR-only and rule-based approaches. A combination of text-based, contextual, and layout-aware methods provided by the methodology, created a scalable, precise, and ethically viable automated invoice data extraction system.

BERT

BERT was taken as a powerful base model in token-level classification. The entity tags indicating the label of each token in the invoice text were B-vendor, I-vendor, B-date, or B-total. BERT bidirectional attention mechanism helped it to acquire contextual relations where field labels might be found in various places and made of different words. The tokens were classified after which post-processing was done to transform the predictions into a structured JSON format so that they could be incorporated downstream. BERT was used to measure the performance of the other models because it was accurate at processing semantic relationships.

DistilBERT

The variant that was used is DistilBERT, which is a compressed version of BERT used to process invoices at scale with high efficiency. It was smaller yet still retained the accuracy of BERT but at a faster inference and less resource usage. Similar to BERT, DistilBERT was utilized to classify tokens with tags of an invoice-specific entity. The obtained outputs were formatted in the same standardized research format under the name of the JSON, making it uniform throughout the pipeline. DistilBERT showed that good quality results were possible in resource limited settings, and as such, it would make it especially well suited to real-time applications.

LayoutLMv3

The most recent model in the pipeline was layoutlmv3 that combined textual, spatial, and visual features. LayoutLMv3, unlike text-based models only analyzed the content of tokens, it also used positional relationships in the invoice. This helped it to identify the structural cues, like tables, headers, and aligned amounts, that are typical in complicated financial documents. Using semantic and layout features, LayoutLMv3 was able to achieve much higher accuracy at extraction on different invoice templates. The outputs

were organized as the other models under the same JSON schema, which shows how layout awareness is effective in improving real-life performance of document processing.

3.1.3 Functional and Nonfunctional Requirements

The presented system of invoice data extraction has to meet a set of requirements so that it can be both technical and ethically responsible. These requirements can be divided into two groups: formal requirements which stipulate the particular features and functions the system has to accomplish and non-functional requirements which characterize the quality features of how the system has to perform and offer value. Combined, these requirements will guarantee the solution is strong, reliable, and meets organizational requirements.

Functional Requirements

- Flexibility in Invoices Reception and Input.

The system should be able to accept invoices of different formats in terms of PDF, scanned images, and machine-readable text files. This is necessary due to the flexibility of the fact that organizations get invoices of different kinds of vendors, each of which has a different design, structure, and form.

- Integration of preprocessing and OCR Preprocessing / OCR.

In the case of image-based or scanned invoices, the system should include preprocessing procedures, which include noisy parts removal, skew correction and resolution optimization. The use of Optical Character Recognition (OCR) is recommended to help turn image-based data into readable text to enable the standardization of all invoices into downstream processing.

- LLM Information Extraction.

The system should employ transformer-based models in order to isolate key invoice fields such as the name of the vendor, invoice ID, invoice date, currency, line items and sub totals, tax amounts and totals. The models should be able to generalize on many layouts without the use of strict, template-based rules.

- Adaptability and Flexibility.

It should be capable of including the zero-shot and few-shot prompting methods so that the system can take care of previously unknown types of invoices without having to undergo substantial retraining or use of large volumes of annotated data. This flexibility provides real-life flexibility.

- JSON-Based Structuring

Fields extracted have to be converted to a common JSON schema, and interoperability with enterprise resource planning (ERP) systems, auditing tools, and other downstream apps need to be ensured. The usability and validation of outputs are also enhanced by consistent structuring of outputs.

- Validation and Error Management.

The system should be able to authenticate extracted values using logical checks like date format, subtotals and tax values adding up to the total. To ensure any inconsistencies, the system must raise warning or back-up measures to either review or correct.

- Evaluation and Reporting

The system is expected to produce evaluation reports which are based on the performance metrics of accuracy, precision, recall, and F1-score. These reports will inform the stakeholders on how the systems are performing and areas that need additional improvement.

Non-Functional Requirements

- Accuracy and Reliability

In order to work in practice, the system will have to have a field-level accuracy of over 90%. It should not change its predictions with repetition of similar input conditions.

- Scalability

The system should be in a position to process massive amount of invoices effectively and be applicable to both small businesses and large organizations with high rates of transactions.

- Security and Privacy

Given the fact that invoices include confidential financial and personal data, the system should meet the high standards of data protection and privacy. It cannot afford to hold redundant data and must be certain that all the products are within ethical standards of data management.

- Explainability and Transparency.

Outputs that are extracted should be interpretable and in a clear traceable way. All fields of a JSON are supposed to be translated to corresponding fields of an invoice, and decisions are supposed to be traced where possible to assist in supporting auditing and regulatory compliance.

- Interoperability

The system will need to be designed in a standardized JSON format to be easily integrated with numerous financial systems and ERP systems, which will decrease the cost of integrating the system into organizations.

- Usability

Outputs of the system must be decipherable and comprehensible to both the developers and none technical stakeholders like auditors, financial managers to facilitate wider acceptance by the organization.

- Efficiency

Processes have to be run with a reasonable amount of time even on very large datasets, such that invoice processes are not interrupted by time delays.

- Ethical Compliance

In addition to technical performance, the system should be fair and not prejudiced to specific invoice templates, vendors or languages. It should be developed and deployed in a responsible way that is in line with the practice of responsible AI and ethical standards in the processing of financial data.

The proposed system will be both ethically correct and technically sound by integrating both the non-functional and functional requirements. The functional requirements ensure proper extraction of invoice data whereas the non-functional ensure that the process is secure, scalable, transparent, efficient and fair. The combination of requirements preconditions a solid background of a practical and credible invoice data extraction solution.

3.1.4 Data Flow Diagram

The Data Flow Diagram presents the picture on the way data flows between raw invoice images and formed, validated outputs. Although the high-level context diagram illustrates the relationship between the user and the system, this level of detail describes the in-house modules and repositories to support the entire extraction pipeline. Its main task is to present the logical flow of data involved in the input stage, processing units, language models, and output repositories.

The invoice images are first uploaded by the user or organization into the system. All the data in this project will be derived in FATURA dataset, which will only comprise of invoice image files and their respective Original_JSON annotations. The images are the raw input and reflect a picture of invoices in the real world where there is diversity in terms of vendors, layouts, languages, and currencies. After loading, the invoices are then sent to the pre processing module.

The preprocessing phase takes care of pre-processing the invoices to the next stage tasks. In the case of image based data, noise elimination and skew correction algorithms are used to make the data readable. The invoice images are then converted into machine-readable text with the use of Optical Character Recognition (OCR). Together with this, semantic normalization is made such that, dates are put in the ISO-standard format, currencies coded into a common three-letter abbreviation, and monetary values are formatted in a standardized format of decimal representation. This is where the ground truth annotations of the FATURA Original_JSON files are also read in and matched with the actual invoices. This step gives two parallel outputs; a normalized text stream of text-based models and an image-text representation of layout-aware models.

After preprocessing, the data is directed into the model application module where three

complementary transformer-based architectures are used. DistilBERT is a lightweight system that gets the normalized text and extracts the target fields effectively which can be a significant solution in terms of processing large quantities of data. The text of the OCR is also subject to BERT, which provides a more in-depth semantic interpretation and better contextual forecasting. Meanwhile, LayoutLMv3 works with the textual content and the spatial layout of the invoice and incorporates positional embeddings based on the bounding boxes in order to extract structural information, including line-item tables and aligned totals. All the models produce the forecasts of the four primary invoice fields of the invoice date, currency, subtotal, and total.

The predictions are sent over to the post-processing and structuring module which converts raw model outputs into a standardized JSON schema. This schema provides the interoperability of this schema with enterprise resource planning (ERP) systems and auditing tools. This is the step where validation is carried out such as schema, arithmetic verification of subtotals and totals, and empty or sanity verification of outputs. In case of inconsistencies, fall back mechanisms go off, fall back mechanisms can include processing with a different model or using heuristic fixes.

The validated data in form of structured data is then forwarded to the evaluation and reporting unit. In this case, the predictions are compared to the FATURA ground truth annotations to the held-out test set. Such performance measures as accuracy, precision, recall, and F1-score are calculated to evaluate the effectiveness of models. Confusion matrixes are used to indicate frequent misclassifications, like subtotal and total, and summary statistics provide an overall system reliability. The evaluation reports are then stored and reported back to the user in an interpretable format which gives a numerical performance report as well as a visual representation of the same.
Internal repositories are also kept throughout this pipeline to facilitate traceability and modularity. These are image store of raw invoice images, text store of OCR results, ground truth store of JSON annotations, and a prediction store of model results. The system also produces a flow of data through these repositories in an organized manner, which will ensure that all the stages here are auditable, reproducible and can be extended in subsequent deployments.

In a nutshell, raw invoice images and Original_JSON annotations are processed through preprocessing which includes normalization and OCR for extraction, through three complementary transformer models, and finally give structured JSON outputs which are checked against ground truth. The design ensures effectiveness, precision, and interpretability which makes the system sound both in scholarly research and application in financial records processing.

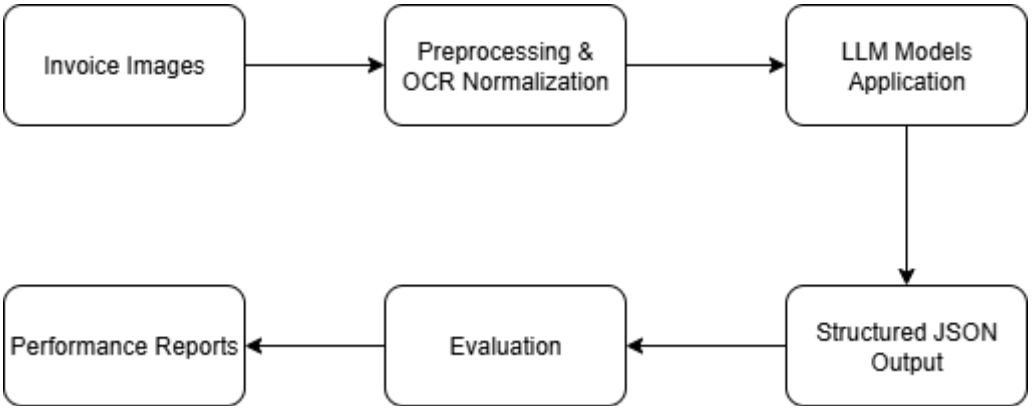


Figure 3.2: Data flow diagram

3.2 Detailed Methodology and Design

In this section, the design of the end-to-end information-extraction pipeline related to invoices is described, and divided into four closely interconnected phases: dataset collection, preprocessing, model application, and evaluation. The design decisions about the design were made in such a way that they have provided methodological clarity, reproducibility and robustness in the face of heterous invoice layouts without compromising the real implementation options that are employed in this study.

1. Dataset collection (FATURA)

The current research utilizes the FATURA dataset as the only source of data. FATURA has only two types of artifacts, namely: (i) the invoice image files (raster images) and (ii) the Original JSON annotations, described as human-curated field values (and, where necessary, region coordinates) of each image.

The dataset consists of invoices of different vendors and layout templates, with great diversity in the level of typography, table format, language clues, and currency conventions. This variety is necessary to test layout sensitivity and language strength.

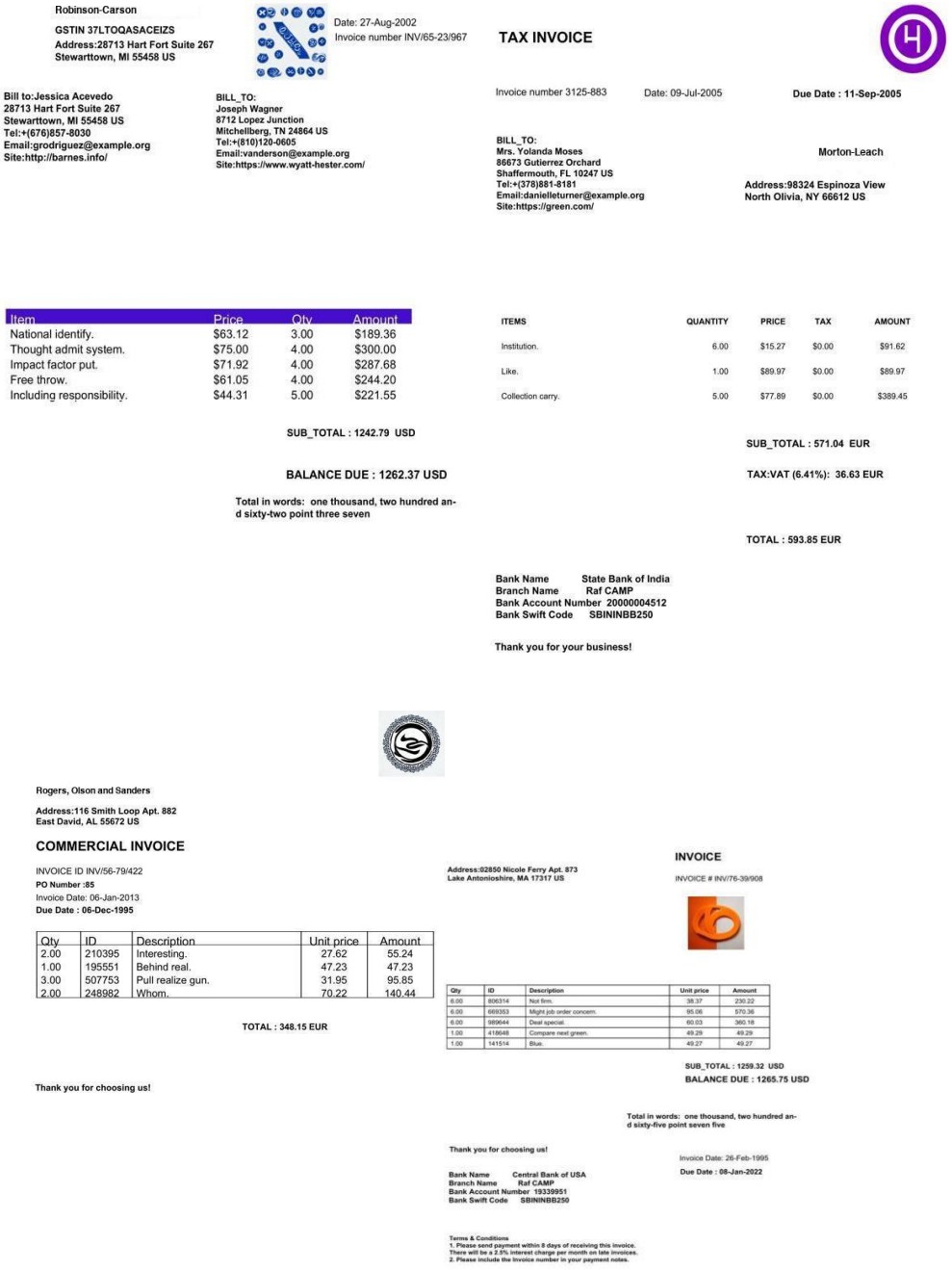


Figure 3.3: Invoice Dataset

2. Preprocessing

Preprocessing takes the raw images and ground-truth annotations and transforms them into harmonized inputs to be fed into downstream models and evaluation.

Light is enhanced to images to reduce scan artifacts in a typical office capture (e.g. denoising, deskewing and resolution normalization). Such operations provide the stability of text legibility and downstream variance without any changes in the semantic content of the documents.

Since the two out of three models are text-based, optical character recognition is used to get a readable machine-text version of each page of the invoice. Coarse line structure is also maintained by the OCR step, and enhances subsequent question-answering at long context.

Extracted date, currencies, and numeric amount candidates are normalized to be commensurably compared across vendors and styles before evaluation as well as when processing model results:

- Dates can be understood with slack to various different formats and as ISO-8601 (YYYY-MM-DD).

- The currency indicators are normalized to an uppercase 3-letter code (e.g. TRY, USD, EUR) and regular symbol-code mappings are used where necessary.

- Canonicalization of monetary values occurs by eliminating locale-dependent grouping characters and equalizing separators of decimals.

Based on the FATURA original JSON files, a simplistic evaluation schema is built based on four fields that are regular across templates and are the key to accounting workflows; invoice date, currency, subtotal amount, and total amount. This schema is mapped to the original annotations and quality checks are made to ensure that there is a one to one mapping between each image and its normalized label record. These are the fields that are not under experiment (e.g. line items or document identifiers) to maintain the task specification accurate.

3. Model application

They use three models that are based on transformers and have a complementary inductive bias in the text and page structure. In order to maintain comparability, the output of all the models would be in the same normalized schema as outlined above.

DistilBERT (question answering based on text). An extractive question-answering paradigm is applied on the OCR text using a compressed transformer. All target fields are elicited using a well thought-out natural-language prompt (e.g. the date of invoice as a date only or the total as a number only) and subsequently lightweight post-processing transforms the span chosen into its normalized form. DistilBERT is focused on computational performance and throughput with a high level of general semantic capacity.

BERT (text based question answering). The same question-answering formulation is applied on the OCR text by use of a full-size bidirectional transformer. Compared to the compact model, this model offers more contextual representations, enhancing disambiguation of linguistically difficult invoices (e.g. unconventional labels on the

vendor, mixed-language tokens or unusual date construction) at the expense of increased compute.

LayoutLMv3 (question answering multimodal document). Layout-aware transformer takes the invoice image as input in a fashion that shares textual information, space utilisation and visual hierarchy. Document level prompts are made, but predictions come as a result of multimodal attention between the content and the layout. It is especially good in cases where there is a need to differentiate closely related monetary fields (e.g. subtotal vs. total) and where line-item tables or alignment patterns have decisive information that is not immediately apparent in plaintext.

In all models, inference produces raw answers with, where there is model confidence scores, model confidence scores. Few deterministic guards are used to minimize shared confusions (such as the suppression of a prediction that is absolutely identical to the prediction of the subtotal, with the exception that lexical cues have confirmed it). The resultant values are again normalized in the same date, currency and amount rules used in the preprocessing process such that the resulting semantics across models are the same.

4. Evaluation

The principle according to which the assessment plan is made is that the discrepancy in modeling must be assessed at the same level with the same targets and equalization.

Targets and serialization:
The comparisons of predictions and ground truth occur on the four fields (invoicedate, currency, subtotalamount, total amount) on each invoice of the held-out test partition. They are both first processed by the common normalization layer to eliminate format artefacts as a source of spurious error. Lightweight record Predicted ends are coded in lightweight record format to allow deterministic repeatable scoring

Other possible alternatives to be considered:

Prior to the end product of the existing pipeline, a number of alternative invoice data extraction solutions were taken into consideration. The first alternatives included traditional approaches that were based on rule-based OCR and regular expressions. These algorithms read OCR results directly with hand-written templates or regular expressions in order to locate important fields (such as totals or dates). Although this was easy to apply to small and controlled datasets, it was extremely sensitive: any small change in the layout of producers, fonts, or formats would trigger parsing errors. The ongoing manual revisions of these systems to accommodate each new template prevented them using them on a dataset as varied as FATURA.

The other alternative involved using template-matching and heuristic, in which the invoices are classified into familiar layout shapes and are handled under a set of preset positional rules. In the case of well-documented templates, this method can result in precise results. But practically invoices are often not laid out according to expectation, and often contain foreign languages or currencies mixes, as well as being poorly scanned. Template-based heuristics do not work at all in these situations, which restricts scalability and generalization.

End-to-end computer vision pipelines were also taken into consideration. These methods circumvent OCR and seek to identify fields directly by images, based on object detection models or object segmentation models. These methods also have the advantage that they need large quantities of pixel-level annotated data, which were not available to FATURA. They are also likely to give poor performance on documents with a lot of text compared to language models that are able to utilize semantic context.

Having considered these options, it was decided to use a pipeline based on Large Language Model (LLM) with three complementing architectures, which are DistilBERT, BERT, and LayoutLMv3. This option had a number of benefits:

- Semantic robustness: In contrast to regex or templates, LLMs understand contextual meaning and can draw the line between similar fields (e.g. between the number of the invoice and the number of the reference).
- Layout awareness: LayoutLMv3 has spatial structure, which allows correct interpretation of totals, subtotals and line items in any invoice layout.
- Efficiency and scalability: DistilBERT is a lightweight concept to scale in terms of high-volume processing; BERT and LayoutLMv3 can be more accurate in case of sufficient resources.
- Less manual overhead: The models are not specific to visible templates but instead generalize over unseen templates generating lower maintenance costs.

The end design was therefore not only selected because it has greater accuracy of extraction but because it was able to compromise on performance, scalability, and ethical compliance unlike the alternate solutions.

3.3 Project Plan

In our thesis project, we have presented a detailed plan that will see the project completed effectively and in time. Our starting point was a research and requirement analysis stage, during which we researched available literature about invoice data extraction, OCR, and Large Language Models (LLMs), and other ethical considerations, including data protection and compliance with the laws. The next step was associated with data collection and preparation where we gathered numerous samples of invoices and cleaned and anonymized them to then train models. Then, it was time to proceed to the model

building step, during which we added OCR to extract text and refined an LLM to recognize and extract useful fields on invoices, producing a structured output in the form of a JSON file. Once the model worked correctly, we proceeded to system integration and implementation, which is the process of integrating all the components into a single prototype and adding functionality to interact with users, provide security and support legal requirements. Another operational activity we performed during the testing and evaluation phase was to test our system with regard to accuracy, efficiency, and ethical alignment using a number of test methods. Lastly, we tabulated our procedures, findings and observations and thus were able to prepare our final report and project presentation. This strategy allowed us to divide the roles, work together, and meet our research objectives systematically and effectively. In our thesis project, we have developed a viable and efficient plan to deliver effective completion on time. Our initial stage was a research and requirement analysis period during which we researched the available literature on invoice data mining, OCR, and Large Language Models (LLMs) and ethical considerations, including data protection and regulatory compliance regulations. The next step involved the data collection and preparation where we gathered a set of invoice samples and cleaned and anonymized them in preparation to use them in training the model. Then, we began a model building step and combined OCR tools to extract text and fine-tuned an LLM to extract and identify the required fields of an invoice and provide them in a structured JSON format. Once the model was functional we proceeded to system integration and implementation where all components are integrated into a complete prototype and functionality added to allow the user interaction, security and compliance with laws. In the process of testing and evaluation, we were able to test the accuracy, efficiency, and ethical appropriateness of our system with a mixture of test techniques. And, lastly we recorded our procedures, findings and thoughts and this led to the creation of our final report and project presentation. This strategy helped us to divide roles, work as teammates, and accomplish our research objectives in an orderly way.

3.4 Task Allocation

- Shahnur Islam Bishal:

  In charge of the literature review, finding ethical and legal issues and writing the methodology and background part. Additionally does documentation, drafting of reports and assists in designing of the ethical structure of data extraction.

- Sabera Ryhana Mayesha:

  Responsible in data collection and preprocessing (OCR, anonymization), fine-tuning and deploy of the Large Language Model on invoice extraction, and in formatting the output in a JSON structure. As well tests and evaluates and integrates the model and real-world invoice data.

3.5 Summary

This part is dedicated to the creation of an ethical invoice data extraction system based on Large Language Models (LLMs). The technique starts by collecting a wide range of invoice papers that has numerous sources so that the data represented is in a variety of

forms, languages, and formats. This variety enables the successful generalization of the algorithm to see real-world invoices. After the scanning, the invoices are converted to machine-readable text by Optical Character Recognition (OCR) to provide a formal structure to analyze the invoices further. The text is then extracted into a fine-tuned LLM specifically trained to identify and extract relevant fields in invoices including invoice numbers, dates, vendor details, line items and total amount. The outputs of the model are then organized into conventional JSON formats and can be easily incorporated into the accounting or enterprise resource planning (ERP) systems. This does not only make it easy to enter data, but also removes human error and enhances efficiency of processing. Ethical systems are at the core of system design. All of the data processing is performed using privacy-preserving technologies, and the sensitive information is encrypted to adhere to the data protection regulations. Also, the team adheres to strict regulations to make sure that no personally identifiable data are abused or leaked when training models and deploying them. The process of development is divided into several steps. First, the research and requirement gathering define the goals and limitations of the project. This is then preceded by collection and pre-processing of the dataset, which forces the outputs of the model training to be of good quality. The trained LLM is further tested and then ethical and security controls are implemented so as to preserve sensitive data. Lastly, the system is assessed carefully and the outcomes are recorded. The tasks are distributed to the other team members to enable them to proceed side-by-side, whereby one works on technological development and model implementation and the other on research, assessment, and ethical compliance. This is a partnership strategy that offers a compromise between innovation and proper AI practices.

Chapter 4

Implementation and Results

The chapter gives the setting of the environment, testing process, and outcomes of the invoice information extraction system. The models were implemented on Google Colab in Python, OCR software, and Hugging Face Transformers. Standard metrics were used to test the unseen invoices in order to make fair comparisons. Results demonstrate a unique set of benefits of DistilBERT, BERT, andLayoutLMv3 and each of them excels in an area. The results show that the best practical solutions in the real-world application of invoice automation can be hybrid strategies or ensembles.

4.1 Environment Setup

The experiments were based on the environment configuration where reproducibility and uniform evaluation are guaranteed. All the implementations were done on Google colab which was selected due to its ease of use, integration with Google drive and already installed toolchains of deep-learning. Despite the fact that free tier on Colab does not assure the user of access to this GPUs permanently, the use of it as a substitute of the GPUs occurred whenever there was availability, particularly in the cases of the more intensive models like LayoutLMv3. DistilBERT and BERT were the lighter text-based architectures that could be executed on CPU with minimal slowdowns, but when a GPU is present, inference time was significantly reduced by using the acceleration of this architecture.

Python 3.10, Hugging Face Transformers (using Python to load and make inferences on models), Python as the deep-learning framework, and scikit-learn (to measure and plot) were used as programming tools and libraries. Other libraries such as PIL (Python Imaging Library) and PyTesseract were applied in the preprocessing of Optical Character Recognition (OCR). OCR was intentionally noisy to simulate invoice conditions of the real world like fading print, text distortion and stamps. This was so that operational challenges and not the idealised laboratory inputs would have been captured in the evaluation. There was also the implementation of file converters capable of dealing with a variety of formats (.jpg, .png, .jpeg) and the fallback logic dealt with missing or unconventional extensions.

The sample was a set of invoice images with organized ground truth labels in the form of JSONL. The fields of invoice date, currency, subtotal amount and total amount were present in each of the annotations. Every invoice was tested with all invoices following the same preprocessing and normalization pipeline (OCR) as in training. Here are a few

examples of how fields were standardized: date fields were converted to YYYY-MM-DD to enable the proper matching of fields at the string level. This high consistency provided the results were not biased by cleaning test cases ad-hoc.

Measurements of evaluation were the accuracy, precision, recall, and F1-score in each of the fields. The confusion matrices were created to display the rates of success and failure based on the classes in a visual way especially the ability of the models to either give false positives (incorrect guess) or false negatives (missed prediction). Probability scores were also computed, and receiver operating characteristic (ROC) curves and their area under the curve (AUC) were calculated. These offered more insight on quality and threshold sensitivity ranking, in addition to raw categorical outputs.

Notably, the three models were put to test on the same conditions. DistilBERT and BERT were based solely on OCR-reading and responded to schema-based prompts in a question-answer way. LayoutLMv3, conversely, combined text and positional embeddings giving it the ability to use layout information that the text baselines could not have access to. This arrangement not only provided equality in comparisons, but also gave us the ability to identify areas in which architectural differences had a direct effect on strengths and weaknesses.

4.2 Testing and Evaluation

The testing and evaluation stage was aimed at determining the extent to which the models, DistilBERT, BERT and LayoutLMv3, could extrapolate to unknown invoices in the real world. All the models were planned to the identical experiment, preprocessing pipeline, and evaluation metrics to make sure that the experiment was fair. It aimed at coming up with a field level comparative study that would not only score the individual architectures numerically with regards to performance but will also demonstrate the strengths and weaknesses of each architecture.

Testing Procedure

- The test sample of invoices contained two items within each invoice.
- Image of invoice in a normal form.
- Available ground truth annotation in the form of a JSON file of the following key fields invoice date, currency, subtotal amount, and total amount.

All the invoice images already passed through the OCR pipeline, and the PyTesseract was used, which could extract the information about the position and text. For text-only models (DistilBERT and BERT), the OCR text was the input text, and served as input context to question-answering interface. Each schema field (e.g.) had natural language prompts made on it. In what date/period is the invoice?), and the forecasted model span was vis-a-vis the ground truth. In the example of Layoutlmv3, positional embeddings

were provided to text OCR and the model was at liberty to use the content of the text and the arrangement of space in its document question-answering strategy.

To avoid the possibilities of memorization, the test set was quite distinctly separated with the training data. This division was required to ensure that the performance was founded on real generalization competency, and not information leakage. In order to maintain preprocessing consistency, the results were normalized: the dates were converted to universal YYYY-MM-DD format, where it was needed the fields with numbers were rounded, and the extra punctuation was eliminated. The given measures made it possible to avoid the insignificant discrepancy of the findings (e.g. 12/01/23 vs. 2023-12-01).

Evaluation Metrics

A set of highly developed measures of classification was used to measure performance:

- Accuracy: the overall percentage of the accurate predictions compared to the number of cases.

- Precision: the rate of the model predictions that have been correct (resistance against false positives).

- Recall: The test properties (resistance to false negative) are the percentage of true labels to be limited.

- F1 Score: The harmonic mean of the precision and the recall that provides a balanced view especially when either of the two is much less than the other.

These measurements were then calculated on a field by field basis and Macro/micro averages of each field. The macro average treated all the fields as equal entities and the micro average treated all the fields as proportional to its occurrence.

Further, confusion matrices were applied to each of the fields to provide a disaggregation of the success and failure to make a correct prediction which can be interpreted. These visualizations showed either systematic over- or under-projection of specific values by models, failure to represent entire fields, or similar values (e.g. subtotal and total).

To explore further the problem of confidence calibration, we have also constructed Receiver Operating Characteristic (ROC) curves and have also computed the Area Under the Curve (AUC) when probability scores were available to fields. ROC curves indicated model ability to distinguish between positive and negative cases at a number of different thresholds. Close values of 1.0 of AUC indicated good ranking capability but poor categorical performance, but lesser values indicated poor separability.

Evaluation Consistency

The following were the conditions under which all the three models were tested:

- All the images of the invoices were preprocessed with identical preprocessing routines.

- The metrics were measured using the same assessment (classification report confusion-matrix ROC curve analysis).

- Comparison has been performed on the same ground truth labels and not a preference has been given to any model which can be formatted and normalized in some way.

This kind of regular testing procedure enabled the making of the performance differences solely based on the architectural merits and demerits rather than the preprocessing or assessment variation. Since the same conditions were observed, the evaluation facilitated pointing to the actual comparative advantage of DistilBERT, BERT, and LayoutLMv3 that prerequisites the description of the results and discussion in the following section.

4.3 Results and Discussion

The testing stage indicated that there was a great difference among the three models. Although they were all meant to extract the same invoice fields each of them had the bias of their architecture and representational strengths that led to their specialization in different fields. Discussion of the results per model is presented below and finally, a cross-model comparative analysis.

Being the lightest and the smallest model, DistilBERT set a benchmark in terms of performance.

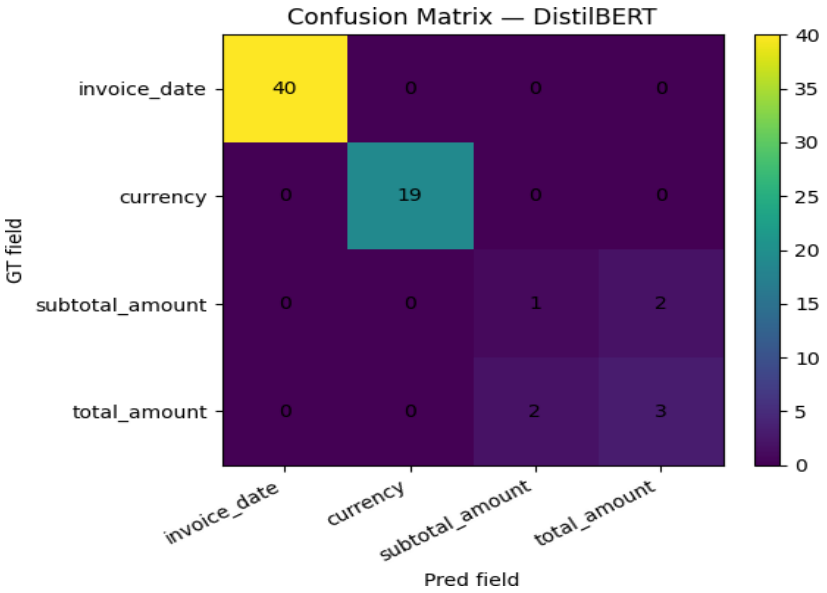


Figure: 4.1 Confusion Matrix DistilBERT

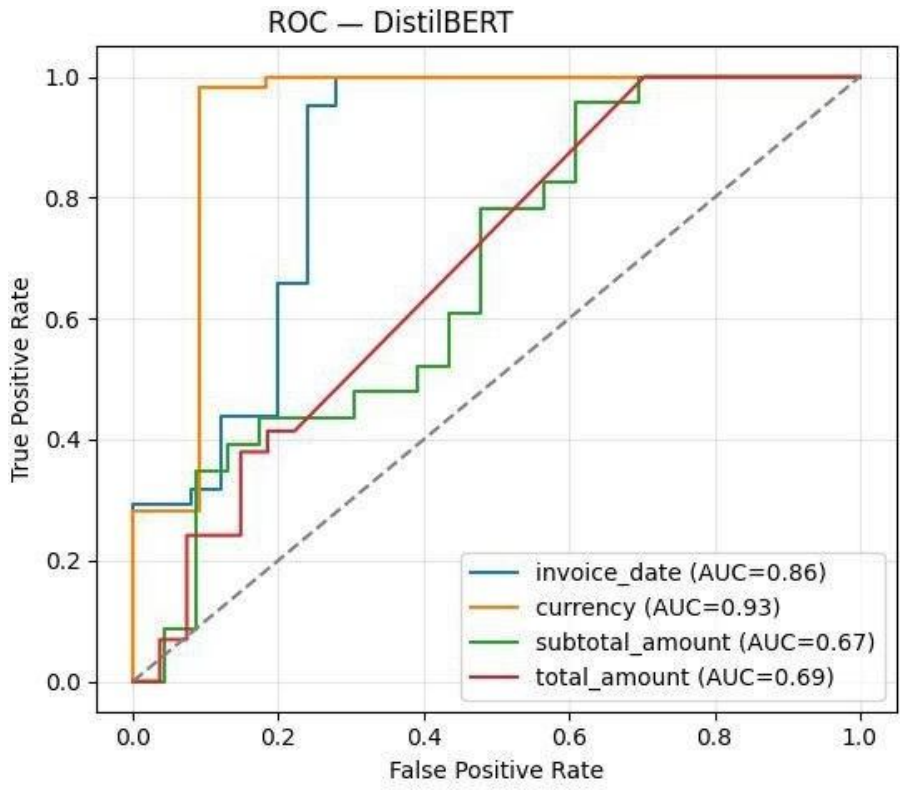


Figure: 4.2 ROC Curve DistilBERT

Its strongest field was invoice date. The model had a precision of 0.889 and recall of 0.606 giving an F1-score of 0.721. The implication of this is that the model was able to reliably extract dates, but it sometimes failed to do so when there was OCR distortion. The confusion table verified that, although the majority of the cases were appropriately identified, close to 40 percent were false negatives.

The currency had worse performance with a weaker F1-score of 0.388. Although ROC curve of currency had high AUC of 0.93, the model had a poor actual recall (0.297). This indicates that DistilBERT might rank probable currency tokens in the right order but not to make predictions in numerous instances contributing to under-detection.

Subtotal amount: Subtotal amount was the weakest, and the F1-scores of F1-subtotal amount are nearly equal to zero (0.034 and 0.081 respectively). In this case, perplexation matrices indicated almost total failure, particularly with numeric fields which were particularly difficult when the subtotal and total values were similar in size or when OCR created tiny artifacts.

All in all, DistilBERT had a macro-average F1 of 0.306 and a micro-average F1 of 0.370. Although not powerful to stand on its own, its relatively decent performance on invoice_date makes it evident that lightweight QA models could work well with textual fields which are well-labeled.

BERT baseline also offered the most equalised performance across disciplines.

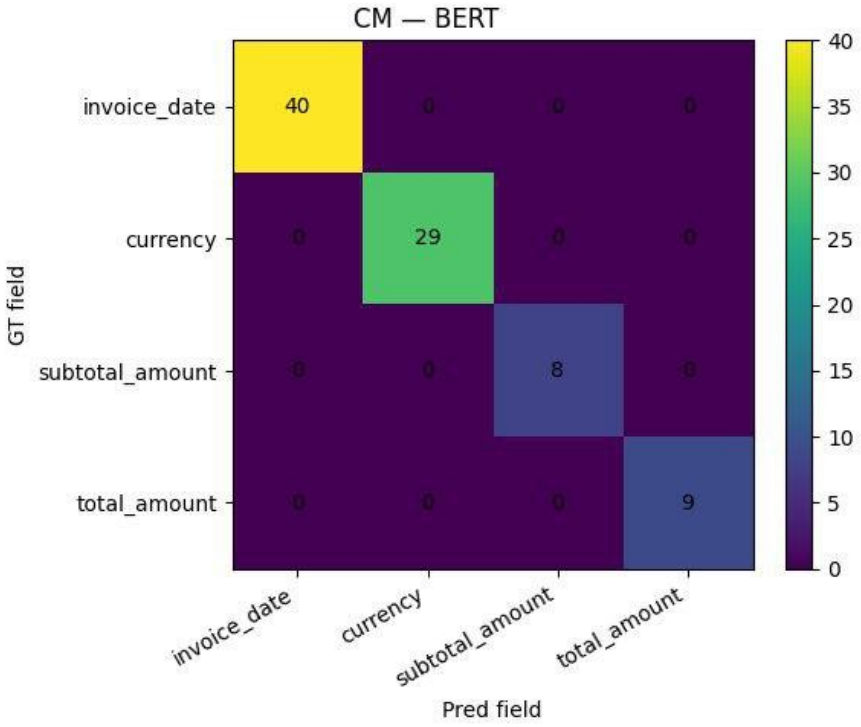


Figure: 4.3 Confusion Matrix BERT

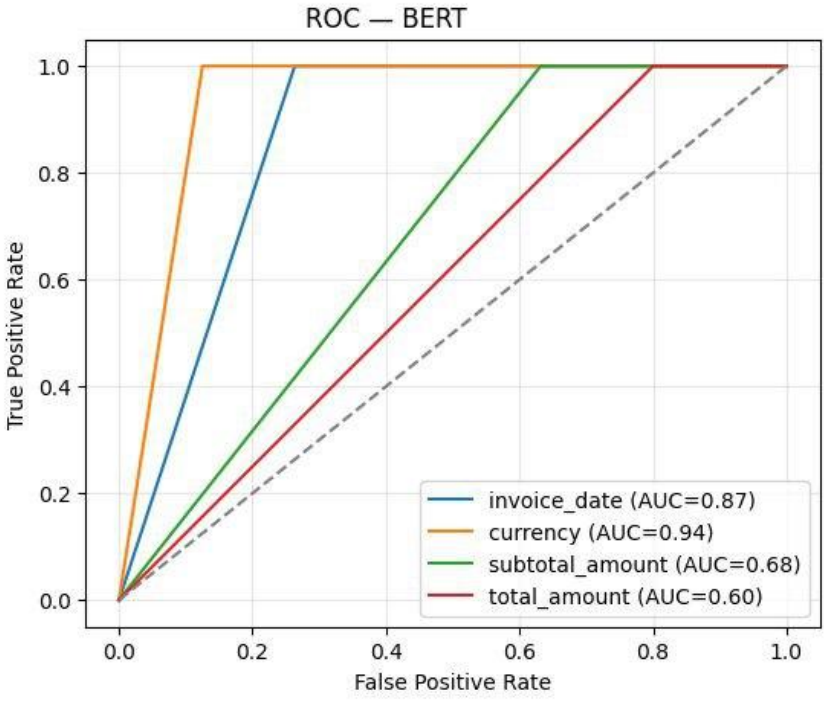


Figure: 4.4 ROC Curve BERT

Invoice_date was treated with the F1-score of 0.708, which is virtually the same as DistilBERT. Accuracy was 0.851 and recall 0.606, once again demonstrating strength but still poor in noisy conditions with a loss of approximately 40% of the dates.

With its F1-score of 0.592 (the highest of all models), BERT was the most successful with respect to currency. Accuracy increased to 0.853 and recall to 0.453, indicating that the model was able to recognize close to half of all the currencies at high confidence. This is due to the fact that BERT has more contextual embeddings and hence it was able to capture currency tokens like USD, EUR, or symbols like the dollar symbol and the euro symbol even when there is noise on OCR images.

Subtotal amount and total amount were weak with F1-score of 0.250 and 0.257 respectively. However, this was a definite step in the right direction as compared to DistilBERT. The confusion matrices indicated that BERT was marginally ready to take the risk of trying to make predictions in these areas which resulted in greater recollection but low precision.

These were further supported by ROC analysis. Invoice_date and currency had a high level of separability (AUC 0.87 and 0.94 respectively). Subtotal and total fields were weaker (AUC 0.68 and 0.60) and the curves showed some opportunities to optimize the threshold. To conclude, BERT was the most balanced model, as it is robust in a variety of disciplines without demonstrating the best in each of them.

LayoutLMv3 presented the evident benefit of using layout information in document understanding.

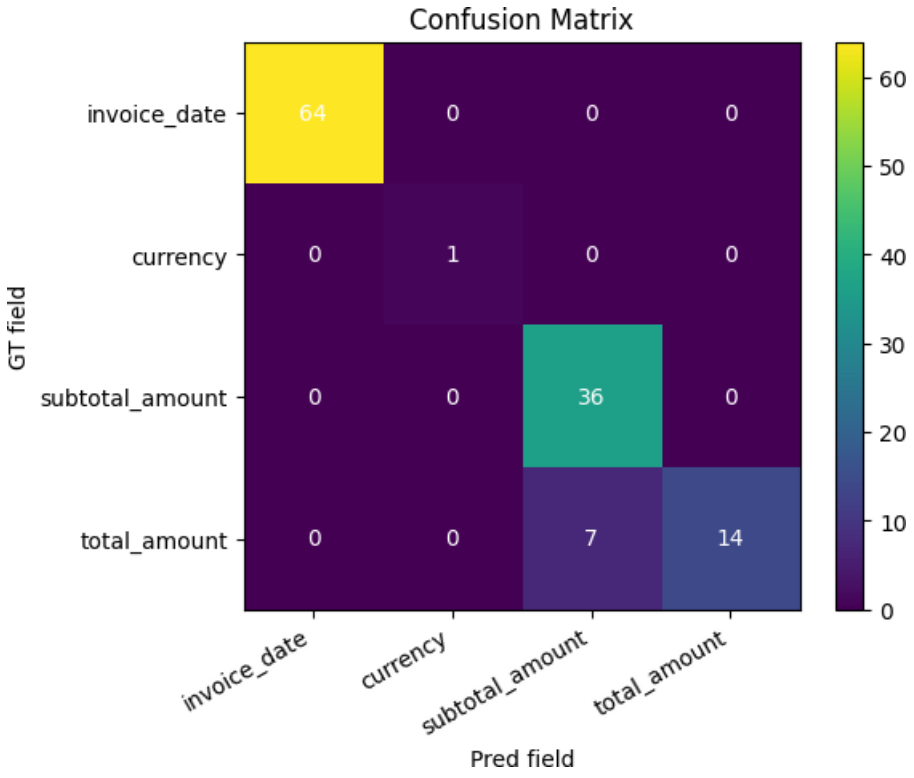


Figure: 4.5 Confusion Matrix LayoutMV3

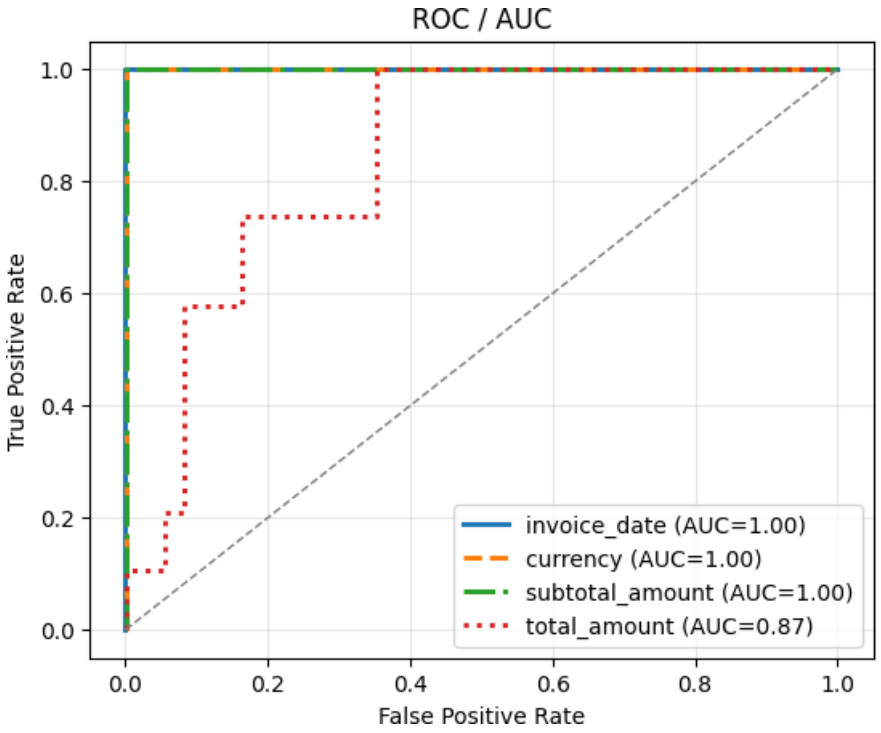


Figure: 4.6 ROC Curve LayoutMV3

The performance of Invoice_date was also almost perfect with a precision of 1.00, a recall of 0.970, and F1-score of 0.985. The confusion matrix showed practically no errors, and it is confirmed that the combination of text and position embeddings provided the LayoutLMv3 with a specific advantage in the areas that are explicitly marked and regularly located.

Subtotal amount also improved significantly where F1 was 0.735. This model was successful in separating subtotals even in cases where totals were physically close and used positional information provided by standard invoice designs.
Total amount was difficult but still better than both text-based models with an F1 of 0.350. The ROC curve showed that AUC of 0.87, which implies that LayoutLMv3 was able to rank totals, but sometimes it confused them with subtotals or taxes.

The most problematic was currency. Though currency ROC curve was perfect (AUC 1.00), the F1 was at 0.031 only. Such inconsistency between ranking quality and measuring recall suggests that there is an extreme thresholding problem: it is only at a few instances that the model predicted money, but when it did it was correct (precision 1.00).

All in all, LayoutLMv3 performed best on structured numeric fields and nearly flawless where it came to date extraction, but its conservative prediction behavior made its performance in terms of currency nearly impossible in practice.

The three models demonstrated expertise areas:

- In the case of invoice date, Layoutlmv3 was the winning model, with close to perfect

results. DistilBERT and BERT were agreeable though restricted by recall.

- On currency, BERT was clear winner, DistilBERT lagged and LayoutLMv3 failed because of over-conservatism.
- In the case of subtotal amount, LayoutLMv3 performed best, then BERT with slight increment and DistilBERT with close to zero performance.
- In terms of total amount, all models performed poorly, of which LayoutLMv3 had the best F1.

Analysis of error in all confusion matrices showed that the type of error that prevailed was false negative. Models were more likely to make no predictions than to risk false positive which would be safer in the financial context but would lower automation effectiveness. ROC curves also showed that all models offered greater ranking ability than categorical results indicated suggesting that threshold calibration and post-processing rules could unlock much performance.

The counterbalance between the models is very impressive. dates, subtotals, currencies, lightweight deployment LayoutLMv3 is the one that is needed, BERT is the one that works best, and DistilBERT is the one that is efficient. An ensemble approach based on a field-specific strategy, that is, having the right model when using a different field, would probably be superior to any single model. Besides, in-field threshold optimization may be useful to reduce recalls, particularly with the predictions of currency of LayoutLMv3.

Limitations comprise errors in OCR (e.g. incorrect recognition of letter O instead of 0), poor performance in narrow numeric fields, and faulty confidence scores. The dataset is diverse and might not represent the entire number of invoice designs that are experienced in actual business. Improvements to be made in the future should be in the form of data augmentation (skewed scans, faded text, added stamps), confidence calibration (e.g., temperature scaling), and hybrid ensemble pipelines. Models that do not use OCR such as LayoutLMv3 are promising, but have to be refined to better deal with small tokens such as currencies.

4.4 Summary

The chapter gave the setup of the environment, the methodology of testing, and the findings of the extraction of invoice information by DistilBERT, BERT, and LayoutLMv3. DistilBERT, which is not very heavy, was the most effective on dates of invoices but weak in numeric features; BERT was the most balanced model, with good results on currency extraction because of stronger contextual embeddings; and LayoutLMv3 was close to perfect in invoice date and also high in results on subtotals, but bad in currency extraction because of its conservative predictions. Confusion matrices indicated that the false negatives were the most commonly occurring type of error and ROC/AUC curve indicated that each of the models had higher ranking capability than the raw categorical output indicated, which showed that threshold tuning and calibration could greatly enhance performance. In spite of the challenges like the OCR artifacts, presence of ambiguous symbols and close spacing of numeric fields, the results proved that the use of transformer-based models can significantly decrease the amount of human effort needed

in processing financial data. Significantly, the complementary advantages of the models indicate that a hybrid model, i.e. using LayoutLMv3 on layout sensitive fields, BERT on currencies, and DistilBERT on lightweight deployments would provide a more reliable outcome in practical invoice automation.

Chapter 5

Engineering Standards and Design Challenges

The engineering standards used in system development, the work's effects on society and the environment, project management with financial analysis, and the mapping of challenging engineering problems are all covered in this chapter. It emphasizes how the suggested invoice extraction system complies with ethical, sustainable, and professional standards.

5.1 Compliance with the Standards

In order to ensure the reliability, security, and interoperability of invoice extraction system, this project is in compliance with the world standards of software, hardware and communications. The choice of standards was also informed by the performance and ethical and compliance issues due to the sensitivity of the invoices that hold confidential and financial data. The use of standardized data formats, such as the use of JSON, cloud-based development environments, and the use of open-source tools were highlighted which will create transparency as well as scalability. A critical evaluation of the options was also conducted before settling on the final design options, weighing their merits and demerits.

5.1.1 Software Standards

The software of the project meets the ISO/IEC 25010:2011 that provides the significant quality attributes such as security, maintainability, usability, and reliability. Since the system must be precise in its outputs, safe with sensitive content and easy to maintain or created to accommodate new vendors and invoice templates, the aforementioned attributes are directly connected to invoice extraction.

All the core models were developed using Python, the widely-known language of artificial intelligence research and engineering. The invoice extraction pipeline was implemented through the support of a video library of machine learning libraries using Python, such as PyTorch, Hugging Face Transformers, NumPy, and Pandas. The code was written according to the PEP 8 style guide to ensure that it is readable, consistent and maintainable. This is especially important as projects that require the combination of multiple factors can be easily complexes, i.e., OCR preprocessing, LLM prompting, and JSON mapping.

The project is based on the JSON schema standard (IETF RFC 8259) as part of structured information. The choice of JSON was because of its native support of REST APIs, human readability and a lightweight format. It also is easily integrated with auditing software

and enterprise resource planning (ERP) systems. Due to its maturity and strong schema validation options, other options such as XML have been considered but was rejected because of its verbosity and inability to be easily read. Proprietary formats were also not allowed as they do not provide interoperability. Contrastingly, JSON makes the most reasonable choice in the real-world financial processes as it maintains the best balance of being readable, concise, and universal.
Some software development alternatives were evaluated. Even though Java and C++ were considered due to their better performance, they did not suit well in the development of the iterative AI model since it was complicated and took a longer time to develop. Similarly, proprietary machine learning platforms were not accepted because they are relatively low in transparency and influence vendor lock-in. Python was the natural solution because of the strong open-source ecosystem and flexibility.

5.1.2 Hardware Standards

Even though the project is centered on software, hardware is also required during the implementation and training of the systems. Processing LLDs is also quite labor intensive with regard to processing invoices, particularly in the case of high-resolution scanned documents. The project conformed to the IEEE 754 standards of floating-point computation in order to ensure similar and consistent numerical operations during the process of inferring and training the model. Such frameworks as PyTorch have this standard turned on automatically to minimize the likelihood of numerical instability.

The project complies also to the ISO /IEC 27001 information security management standards on safe data handling. This will make sure that personal invoice information is coded and stored in a safe location with a restricted number of users. This in real sense means that any type of invoices would be handled as per the universally accepted security models whenever they are uploaded to be tested or evaluated.

There were a number of hardware deployment options that were compared. The on-premise servers are physically secure and fully controlled thus, they were initially thought. They were not so valuable in a project that was mainly intended to research but they were quite expensive to set and maintain. The edge devices NVIDIA Jetson Nano were also tested. Although they were attractive to low-power, real-time, applications, they were not able to run large language models well. Lastly, the cloud-based graphical processing services like AWS and Google Colab Pro were selected to be tested and experimented. These services are affordable, scalable, and meet the requirements of the current standards of data security. Dependence on external suppliers is the trade-off, which is reduced by the fact that the security certifications they comply with are accepted.

The other options like consumer grade laptops or desktop CPUs lacked the power to handle large models that generated long processing time and limited experimentation. It turned out that the cloud GPUs turned out to be the best middle-ground because they are high-performing and versatile and offer standardized environments, which can be both research- and industry-friendly.

5.1.3 Communication Standards

The communication standards are highly valued since invoices most of the time require sending among various components, including uploading documents, sending them to the model and receiving organized outputs in return. In order to ensure the transmission of sensitive invoice information is secure, this project relies on HTTPS with TLS/SSL encryption. The HTTPS protocol is more secure than plain HTTP since it prevents interception and manipulation.

System integration was done by a RESTful API architecture. REST APIs are non-stateful, small, and work well with exchanging JSON data. They facilitate easy connection to ERP software, auditing and cloud platforms. We considered other alternatives, such as SOAP APIs. SOAP is highly typed and has in-built security features, yet is XML based and lengthy making it inapplicable to this case. MQTT also came to our mind since it is efficient in IoT scenarios, but is more efficient with small messages and real-time telemetry, rather than massive invoice payloads. We considered using WebSockets to communicate in real time briefly, but because the system did not require a two-way streaming system, the usage of REST was viewed as simpler and more appropriate.

User access and authentication may be achieved in the system with the help of OAuth 2.0 that is commonly applied as a standard of secure authorization. OAuth will ensure invoice information is only accessed by authorized users. This is particularly significant in a company that involves a lot of roles and authorization that must be handled.

Finally, REST APIs over HTTPS using TLS/SSL were selected since they coincide with the way modern development is operated, and ensure communication is secure, as well as compatible with the JSON output type. Such a mixture is not only easy but also trustworthy and it does not involve any additional effort.

5.2 Impact on Society, Environment and Sustainability

The chapter examines the refrain of the proposed invoice data retrieval system in relation to the larger social objectives, ethical obligations, and long-term sustainability. The design of an information extraction framework based on the JSON format and AI-driven approach has far-reaching implications on the data governance and responsible digital transformation besides businesses and financial stakeholders. The accuracy, transparency, and ethical compliance is the factor in which the system was developed to ensure that automation works to the benefit of the users in a sustainable and trustful manner.

5.2.1 Impact on Life

The proposed system would be of great use in the daily activities of small to medium-sized businesses (SMEs), accountants, and employees. Automation of the time-consuming process of manually extracting and verifying invoice details can save time, reduce monotonous work, and focus on higher value activities such as analysis and decision

making. This will have a direct positive effect on the efficiency, reduced levels of stress, and enhanced productivity in personal and organizational settings.
Moreover, the structured JSON outputs ensure that there is consistency and compatibility across digital platforms and therefore, the system can easily be integrated by users with minimal technical knowledge to their existing workflows. The system enhances confidence of professionals towards financial documentation and helps professionals make decisions easily in a manner that is data driven by reducing human error and simplifying financial record keeping.

5.2.2 Impact on Society & Environment

The project aids in the ongoing digitalization of business and finance activities at the social level. It reduces use of paper-based documentation and provides automated, moral and scalable solution to manage invoices and reduce administrative expenses and encourages environmental friendliness. The change aids in minimizing wastefulness in the organizational procedures and aids in a greater transition toward sustainable digital ecosystems.

The system minimizes unnecessary printing, storage and physical movements of invoices, which is environmental friendly. Digital data extraction also encourages paperless operations which are also in tandem with the international sustainability programs. Also, the system will aid in the minimization of the unnecessary energy use since it can be operated either locally or on the lightweight cloud applications.
In keeping with the Sustainable Development Goals (SDGs) of the United Nations, namely Goal 8 (Decent Work and Economic Growth), Goal 9 (Industry, Innovation, and Infrastructure), and Goal 12 (Responsible Consumption and Production), this system will encourage sustainable business and support the digital equity movement by making the latest and greatest tools available to organizations of all magnitudes.

5.2.3 Ethical Aspects

In this project, the concept of ethical responsibility has been a design principle throughout the project. Privacy, security, and fairness were upheld rigorously since the invoice data is often sensitive containing financial and personal details. All the datasets were anonymized, and no personal or confidential data were stored beyond the scope of the study.

The system gives preference to transparency by ensuring that the results of extraction can be interpretable and verifiable by a human user in cases where it is feasible to do so using interpretable AI methods. This makes people less afraid of automation and reduces the feeling of the black box that is often attributed to AI systems. The consideration of fairness was also considered, which ensured that the model is equally accurate, and not that it discriminates against invoice templates or vendor.

Moreover, the project ensured that only agreed or synthetic invoice data was utilized to

avoid the use of unauthorized data. Every piece of extracted data is preserved against misuse or unauthorized access in the course of deployment due to the optional implementation of secure communication protocols (like HTTPS/TLS). These steps ensure the system operates ethically in real life circumstances by following the principles of accountability, autonomy and non-maleficence.

5.2.4 Sustainability Plan

A complex sustainability plan has been launched to make sure that the system will be useful, relevant, and work in the long-term:

- Technical Sustainability: The modularity ensures Adaptability and maintainsability, and the python and open-source machine learning libraries to which it is built. The system is interoperable with much of the enterprise software and can be integrated to various other systems in the future such as ERP systems or cloud-based accounting software due to the use of JSON-based output.
- Economic Sustainability: The system will be consumer-grade, with the hardware and open-source platform less costly, and thus, small and large business firms will be able to use it. Its automation ensures accuracy and transparency and decreases the labor costs over the long term.
- Environmental Sustainability: The system minimizes the carbon footprint on document handling and document storage by promoting the use of digital finance that is paperless and by minimizing dependence of paper-based processes. Lightweight computation makes the training and deployment procedures energy efficient.
- Community and Knowledge: Under this, the ethical model of the project is safe, transparent and ethical. Standardization, documentation and the possible release into the open source of some of the components are all promoted in a bid to ensure that knowledge is shared and community based development and sustainability are promoted.

5.3 Project Management and Financial Analysis

Project management techniques coupled with good financial analysis played a key role in the success of this project in order to guarantee project completion and applicability over time. The entire project was structured in the following way: the dataset was first developed and marked and then followed by preprocessing, testing the model, analysis, and report. The milestone based planning was able to track all the stages made and follow the progress and make changes. The implementation of agile philosophies involved breaking down of work into small steps and feedback streams to guide the process of refinement at each phase. The process of data storage, model training, and teamwork was conducted with the assistance of Google Drive and free Google Colab tools but version control with the assistance of such tools as GitHub were used to track the changing codebase. Simple planning boards did the task management, which ensured that the testing, final reporting, model fine-tuning deadline and the dataset preparation were met. Such an approach to management was systematic but flexible, and it assisted the project in striking a balance between the needs of the academic research and the objectives of the practice.

The project was carried out in extremely low cost model which utilized free resources where possible, financially. Direct costs on computational infrastructure were nonexistent because the free version of Google Colab was adequate to access GPUs to train and

evaluate smaller-scale models and mid-scale ones. Data set building did not require significant amount of time and money investment since it was predominantly being based on manually gathered invoice images and annotations. There were no requirements to pay a certain sum because such requirements like Hugging Face Transformers, PyTorch, and TensorFlow were open-source and did not cost anything. The overall budget was kept minimal as there was only one direct expense, which was the expenditure on the internet access and subscription to the cloud storage where the final thesis could be stored safely and the printing and binding of the document. This budget would be broken down as follows at the level of research:

Table 5.1: Actual Research Budget (Self-Supported, Research-Based).

Component	Estimated Cost (BDT)	Remarks
Personal Laptop and Devices	0 (existing)	Used for code development and dataset preparation
Google Colab (FreeTier)	0	Free GPU/CPU for model training
Internet and Cloud Storage	3,000	Monthly expenses for uploads, backups, and collaboration
Open-Source Libraries	0	Python, TensorFlow, PyTorch, Hugging Face, etc.
OCR Tools (Tesseract, OpenCV)	0	Fully open-source
Printing and Binding	2,000	Hard copy submission of thesis

Total Research Budget: ~7,000 BDT

A real business application of the system would require more funding to grow the system to scale, provide security and professional usability but the self-sustained academic path was viable. Probably, this would be a software-as-a-service (SaaS) system within an enterprise setting that would require enterprise-level OCR solutions to be more precise, model training and inference to be dedicated to a server, adherence to global data privacy laws, including GDPR and ISO/IEC 27001 and the creation of a user-friendly mobile/web interface. Commercial implementation would also require expenditures on dataset development through expert labeling service, cloud computing, customer support and marketing to users.

Table 5.2: Alternate Budget (Enterprise-Scale Deployment).

Component	Estimated Cost (BDT)	Remarks
Dedicated GPU Cloud Servers	450,000	Three months of enterprise-grade compute

		resources
Professional OCR Software	50,000	Tools such as ABBYY or Google Vision API
Dataset Expansion and Labeling	100,000	Professional annotation services
Compliance and Security Audits	75,000	GDPR, ISO/IEC certification costs
Web/Mobile Application Development	120,000	Full-stack interface for SaaS deployment
Cloud Hosting and Database	80,000	Yearly hosting, maintenance, and storage
Marketing and Awareness	50,000	Campaigns to reach SMEs and accounting firms
Ongoing Support and Maintenance	60,000	Continuous updates and customer service

Total Enterprise Budget: ~935,000 BDT

The difference between academic and commercial viability is stressed by the differences between the two budgets. Whereas the enterprise budget reflects the real expenses of developing the solution into a reliable, safe, and user-friendly service that suits the financial sector, the academic budget reflects how the top-tier AI studies can be conducted at low costs with the help of free platforms and open-source applications.

Budgetary aspects were also planned, as well as the revenue model that would be implemented to ensure the financial sustainability of the system in case of the eventual commercialization of the system. An example is the subscription-based SaaS model, where businesses pay a certain monthly cost depending on the number of invoices that they handle. Simple invoice extraction functionalities would be offered free under a freemium business model, with additional functionality such as bulk processing and integration with ERP being available under higher business tiers. The extraction engine would be able to be embedded into the systems of larger firms and ERP providers through enterprise licensing agreements. A pay-per-use API may also help smaller firms to handle documents on-demand, which will provide them with flexibility and generate a predictable income. By diversifying its revenue model, the system was capable of serving both the small and medium sized business and at the same time be able to expand to cope with the demands of an enterprise.

Overall, proper project management ensured that it delivered goals punctually and effective utilization of resources, and the financial analysis depicted that the academic pathway was cheap and that it was realistic to deploy it on an enterprise scale. These factors together give a very good road map on the long-term commercial potential, practical viability and academic contribution of the system, when considered as a whole.

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

This research project took several multifaceted methods of solving engineering problems to be accomplished successfully from the data preparation and OCR integration to the process of model training, evaluation, and deploying it in a form of a JSON resource. The mapping shown below demonstrates how the work has been aligned with the Engineering Problem (EP) framework.

Table 5.3: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓			✓

Justifications:

- EP1: Level of Knowledge: The project demanded extensive knowledge related to transformers based architectures, including BERT, DistilBERT, LLaMA and Mistral, natural language processing (NLP), OCR, and JSON schema design. The experience in pipeline creation involved learning about the PyTorch/TensorFlow and Python programming and evaluation metrics.
- EP2- Scope of Competing Requirements: The system had to be of high accuracy, but had to be small enough to be run on limited hardware (Google Colab free tier). It had to apply to both the accounting companies and the SMEs and have to meet the ethical requirements regarding data privacy.
- EP3- Depth of Analysis: Confusion, accuracy, and F1-scores among numerous models were all systematically analyzed and their extraction completeness was assessed. Efficiency versus complexity trade-offs also were identified as a result of comparative analysis of models that have been designed to become the most efficient.
- EP4- Issue Familiarity: It was processed and refined and well commented to assist in fixing the anticipated problems, including inconsistencies in invoice layouts, OCR errors, bad quality of noisy images and un-tagged data.
- EP7- Interdependence: It was built around interdependent modules, including JSON schema to standardize it, OCR to extract the text, and LLMs to process the text in a more organized structure. The design interdependence was achieved in terms of interaction of the modules among themselves.

Mapping based on Knowledge Profile:

The profile of knowledge (K1-K8) is also compatible with the project because it requires the background of engineering, professional knowledge, and research integration.

Table 5.4: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
		✓	✓		✓		✓

Justifications:

- K3- Engineering Fundamentals: The classification, optimization and assessment metrics used to determine the model accuracy were some of the applied principles.
- K4- Specialist Knowledge: The data extraction pipeline was created with the help of the specific expertise of NLP, LLMs, and OCR systems.
- K6- Engineering Practice: To obtain repeatability and rigor, the project followed the engineering practices which contain data preprocessing, fine-tuning of the model, systematic evaluation, and reporting.
- K8- Research Literature: A thorough review of the available literature in the area of document AI, OCR, and invoice extraction informed the methodology and ensured that the study addressed the research gaps.

5.4.2 Engineering Activities

The project cycle also reflected a number of sophisticated engineering phases, such as dataset generation, training of the models, testing them, and designing the system. Complex Engineering Activities Mapping.

This section aims at showing how different activities were integrated into the project lifecycle by mapping the entire issue onto the Engineering Activities (EAs). Creative thinking and close observation of the impacts on the society and the environment also required a number of resources, stakeholder engagement, and thinking to come up with an ethical invoice data extraction system. The project is a good example of how engineering practice extends beyond technical implementation to broader effects and professional obligations by aligning itself with EA1-EA5.

Table 5.5: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓			✓

Justifications:

- EA1- Diversity of Resources: Google Colab, GitHub, transformer libraries, and Open

Source OCR tools were utilized in the project. The variety of the resources can be seen in this rich toolkit.

- EA2- Level of Interaction: In creating the systems, a variety of views was taken into account because of peer review, academic supervision, and consideration of the needs of the accountants and SMEs.
- EA5- Familiarity: There was a lack of resources to carry out the project since it involved fast prototyping and experiments, which were feasible only through the previous coursework and experience using the NLP and ML tools.

5.5 Summary

This chapter explained the main principles of engineering and complicated design issues connected with the creation of the proposed system of invoice data extraction with the assistance of large language models. The project needed to be done with a multidisciplinary approach and would combine transformer-based deep learning, optical character recognition (OCR), natural language processing and ethical data governance. All steps of the development process were associated with actual financial applications, including creating a dataset containing annotated invoice images up to creating any form of output based on JSON.

The system was able to comply with the accepted engineering standards with the assistance of open-source libraries (PyTorch, TensorFlow, and Hugging Face Transformers), systematic preprocessing, fine-tuning of the model, and comprehensive testing. Explainability was introduced to resolve the usual predicament of most AI systems that was the blackbox that extract fields to be verified and be trusted. The structure ensures the compatibility of the tools with the existing enterprise tools by generating formatted JSON data, which are easier to work with by SMEs, accountants, and auditors.

The project also dealt with other more general sustainability and ethical issues. It also saved on money by means of free-cloud technology, provided incentives towards paperless digital financial operations to reduce environmental footprint and paid off excessively with regard to privacy when dealing with sensitive invoice information.

This chapter was concluded by showing that this research could comply with engineering, ethical and practical standards. The suggested system will make financial operations more transparent, scalable, and sustainable, as it will offer a scalable AI-driven system that can be easily accessed and interpreted to automate the process of data extraction of invoices.

Chapter 6

Conclusion

The chapter focuses on the importance of effective and ethical invoice data extraction using Large Language Models (LLMs). It shows that with the help of LLMs, one can accommodate different formats of invoices, deliver structured JSON responses, and enhance transparency, security and regulatory adherence at the minimal amount of human labor. It also notes such constraints as the diversity of data, reliance on high quality training data, the cost of computing, privacy concerns, and generalization problems. Multilingual support, real-time processing, and explainability will be the focus of future work and strive to produce powerful, scalable, and ethically accountable designs to address the global financial flows.

6.1 Summary

This paper demonstrates the crucial importance of ethical information retrieval when it comes to the processing of invoices using Large Language Models. Invoices often include highly sensitive financial and personal data and not properly automating their processing may put organizations at risk of data breach, deceit, or governmental nonconformity. This work provides a responsible but technically advanced method of using LLMs which are capable of interpreting the textual as well as the contextual elements in documents. In contrast to traditional OCR-based systems or rule-based extraction, LLMs can be customized to most types of invoice this eliminates mistakes and eliminates the necessity of manual intervention. Stressing on ethical design can help to make AI systems not only effective, but also safe and trustworthy when working with sensitive financial information. The study provides a respectable AI-based solution, which combines the strength of LLMs with format-driven JSON responses. The solution maps extracted invoice field encompassing vendor name, invoice ID, date, subtotal, tax, and total amount to a pre-determined JSON format, whereby the solution offers consistent and machine-readable outputs which can be effortlessly integrated into the accounting or enterprise resource planning (ERP) systems. This is also a structured method that involves the use of automated validation that ensures the presence of numerical accuracy, missing data, and bad formats. Moreover, the fact that JSON outputs are used enhances transparency since the businesses only need to monitor the manner in which data is collected and processed. This openness is crucial to the assessment, regulatory adherence, and the creation of trust in AI-driven financial processes. Scalability and adaptation to the changing laws and the different formats of invoices are also a priority in the project. The pipeline includes ethically issues of AI such as bias mitigation, data security, and the interpretability of the model. Mitigation of bias ensures all data are processed (regardless of vendor, location or type of invoices) in the model and high data security guidelines prevents sensitive data being accessed or compromised. The interpretability of its model

allows financial professionals to find out how decisions are made and validate extraction results, thus making them more responsible. The study, by embracing these ideas, appeals to both effectiveness in practice and nowadays there is an increasing demand to have AI systems that are highly ethical, especially in the sensitive financial fields. Going forward, the outcomes of this paper will be useful in the development of strong and ethically conscious AI systems that can streamline the financial processes whilst focusing on ethical concerns. Other improvement areas that will be made in the future include enhancement of system integration, enhancing model accuracy, and adherence to evolving requirements. This may involve adding finer models in multilingual invoices, use of more powerful LLMs or perform real-time extraction operations in business cases. The framework ensures the establishment of AI solutions that can handle the complex financial documentation safely and efficiently, and in a manner that ensures organizational efficiency and social trust towards automated financial processes by balancing performance and ethical responsibility as well as regulatory compliance.

6.2 Limitation

Automated invoice data extraction with Large Language Models (LLMs) has immense efficiency and flexibility advantages, and a number of deep-seated issues and constraints, which must be dealt with accordingly.

1. Data Format Diversity:
Among others layout, language, fonts and format are among the data formats contained in invoices. Some of the bills are plain and clean, and a few of them include several tables, logos, and stamps, or even written comments. Even extremely powerful Large Language Models (LLMs) could not handle extremely complicated or crafted bills. Field positions, and language (like total due rather than amount payable) and document structure could cause differences in accuracy of extraction unless the model was trained on these patterns. Multimodal techniques, attention to preprocessing, layout-sensitive modeling and even consistency and accuracy of extraction are all required to make the extraction of all types of documents consistent and accurate.

2. The Quality Training Data are relied on:
The quality, size and diversity of the training data is of great importance in the efficacy of the LLM. The extraction of invoices needs annotated data sets with great variation in suppliers, format, and geographical location. Unavailability to such real world information can disrupt in the generalizing of the model that results in error in treating new layouts or content. Besides, inoculation of invoices is a manual procedure that is resource- and time-intensive and therefore limits the nature and size of datasets. The issue is that even the high-quality training information are not sufficient to demonstrate the consistent or the full results despite the advanced LLMs.

3. Computational Resources:
Big LLMs need huge computational resources during both training and inference. To compute invoices on a fine-tuning basis, the model requires a lot of memory usage, a

high-power GPUs, and substantial space to store dataset and model checkpoints. The cost of scaling these models (handling of thousands of invoices per day) can be too costly to the medium sized businesses. Resource constraints may also be imposed on experimentation and hyperparameter tuning which are necessary to make the best extraction accuracy. This dependency creates operational issues in as far as the implementation of the billing systems according to the LLM to a real-life and resource-constrained environment is concerned.

4. Privacy and Legal Risk:
Even good ethical precautions might not suffice to stop privacy and legal exposure to sensitive financial information. The invoices are prepared containing confidential information regarding the client names, the name of the vendors, the price of payment, and the bank account number. Hack or software setup can result in breach of data, regulatory or reputation. The legal regulations on data protection (GDPR or local financial privacy regulations) must be followed, and any company lacking an effective controls system is prone to lawsuits. The system design is highly crucial in terms of ethics, but it will not eliminate residual risk completely.

5. Generalization Limitations:
Zero- or few-shot learning is also effective with LLMs, but new or location-specific templates of invoices that could not be easily dissimilar to training data may be challenging. To give an example, an invoice of a supplier in another country may be written in odd languages or currencies or designs thus misclassifying or not giving information. Generalization problems should be re-trained by retraining or through a reduced number of shots or other rule-based heuristics. Otherwise, having not taken such precautions, the system will not be reliable in extracting out of a collection of global or highly diverse invoices, the correct information.

6.3 Future Work

Multilingual assistance is an inseparable future. The current algorithms to extract invoice can today be applied to the documents that are written in one language only, and frequently fail to operate in case the documents are written in several languages or local traditions are used to write an invoice. The greater the number of languages languages the model can support, the greater the enterprises will be in a position to utilize the same model across the borders without necessarily having to utilize different pipelines or re-train their employees to different languages. This does not just entail a translation of text, but also finding out differences in dates, currencies, numeric pattern and geographical names like Total Due and Amount Payable. The multilingual characteristics will promote flexibility of the system and its strength to the extent that it can accommodate many multinational suppliers and cross-regional banking activities.

Another area that should be enhanced is real-time processing. The huge majority of the modern systems, in particular the large LLM-based systems, are highly computational and execution in batch mode. Reducing the extraction chain to real-time performance can

assist the businesses in processing the invoices as they come in and aid the functioning of the high-throughput environments and accelerate the financial operations. The future directions may be based on such strategies as model distillation, lightweight architectures, as well as inference optimization strategies to minimize the latency without affecting the accuracy. Even though it will be particularly handy to large organizations that get thousands of invoices every day, real-time processing will offer them quicker payment and the ability of more reports and increased operational efficiency.

Finally, there are the areas of concern that can be explainable and interpretable that should be of interest to future studies. The systems, which operate on the basis of LLM, are even referred to as black boxes and it is difficult to know how and why certain areas incorporated in the invoices are picked up by the consumers. The fact that techniques of explaining model forecasts are included can greatly increase user confidence and make the automated financial processes more factual. One can gain an insight into the way that the model arrives at its decisions using visualization, confidence scoring and rule-based explanation overlays. This not only simplifies auditing as well as regulatory compliance, but also makes it easier to find and rectify problems. The greater explainability can turn the system into responsible and reliable and suitable to operate in the ethically sensitive financial areas.

These three directions that also comprise multilingual support, real-time processing, and explainability are the most desirable to add and enhance the invoice extraction systems based on the use of LLM. With such areas of concern, the future research will yield scalable, efficient, flexible and transparent systems which can sustain dynamically changing requirements of the global organizations without compromising trust, compliance and operational excellence.

References

[1] Y. Wei, et al. Robust layout-aware information extraction for visually rich documents with pre-trained language models. *arXiv*, 2005.11017, 2020.

[2] S. Amari, et al. An efficient deep learning-based approach to automating invoice document validation. *arXiv*, 2503.12267, 2025.

[3] T. Amjad, et al. An agentic system with reinforcement-learned subsystem improvements for parsing form-like documents. *arXiv*, 2505.13504, 2025.

[4] X. Huang, et al. CUTIE: Learning to understand documents with convolutional universal text information extractor. *arXiv*, 1903.12363, 2019.

[5] B. Güneş, N. Potti, S. Tata, J. B. Wendt, M. Najork, and J. Xie. Data-efficient information extraction from form-like documents. *KDD Workshop on Data-Efficient Learning*, 2021.

[6] Y. Xu, et al. Representation learning for information extraction from form-like documents. *Proc. ACL*, 58:6495–6505, 2020.

[7] P. Patel. Design and implementation of an OCR-powered pipeline for table extraction from invoices. *arXiv*, 2507.07029, 2025.

[8] S. Bhattacharyya, et al. Information extraction from heterogeneous documents without ground-truth labels. *Proc. WACV*, pp. 1–12, 2025.

[9] Z. Zhang, et al. Generating synthetic invoices via layout-preserving content replacement. *arXiv*, 2508.03754, 2025.

[10] H. Ghosh, et al. Information extraction from multi-layout invoice images using FATURA (OCR-free). *Engineering Applications of Artificial Intelligence*, 138:107817, 2025.

[11] A. Sharma, et al. DocExtractNet: Enhanced document intelligence for receipts and forms. *Information Processing & Management*, 61(4):103619, 2024.

[12] R. Singh, et al. Automated invoice processing: Machine-learning based information extraction. *AI Open*, 4:97–106, 2023.

[13] D. M. Rao, et al. Automated invoice handling method using OCR. *Advances in Intelligent Systems and Computing*, 1174:193–202, 2020.

[14] P. Deenadhayalan. Automated invoice data extraction using image processing. *International Journal for Research in Applied Science and Engineering Technology*, 10(6):1711–1716, 2022.

[15] D. Leon. Extracting information from PDF invoices using deep learning. *B.Sc.*

Thesis, KTH Royal Institute of Technology, pp. 1–60, 2021.

[16] C. Lee, et al. FormNet: Structural encoding beyond sequential modeling in form document information extraction. *arXiv*, 2203.08411, 2022.

[17] L. Cui, Y. Xu, T. Lv, and F. Wei. Document AI: Benchmarks, models and applications. *arXiv*, 2111.08609, 2021.

[18] Y. Xu, et al. LayoutLM: Pre-training of text and layout for document image understanding. *arXiv*, 1912.13318, 2019.

[19] M. Sarkar, M. Aggarwal, A. Jain, H. Gupta, and B. Krishnamurthy. Document structure extraction using prior-based hierarchical semantic segmentation. *arXiv*, 1911.12170, 2019.

[20] K. Silva and T. Silva. A review on document information extraction approaches. *Proc. RANLP Student Research Workshop*, pp. 180–187, 2021.

[21] T. C. Nemtoc and A.-M. Ghiran. Natural language querying of invoice data using RAG and GraphRAG: Leveraging LLMs for financial document insights. *Proc. CAiSE Workshops*, pp. 69–80, 2025.

[22] S. Paliwal, et al. TableNet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. *arXiv*, 2001.01469, 2020.

[23] J. Garud, B. Darak, K. Sarode, K. Pandhare, and A. Nair. Invoice extraction using LLM and OpenAI. *International Journal of Research Publication and Reviews*, 5(4):3862–3866, 2024.

[24] I. Hassle and M. Bardvall. Automating invoice recognition: A comparative study of large language models and OCR/ML technologies. *B.Sc. Thesis, KTH Royal Institute of Technology*, pp. 1–65, 2024.

[25] T. Tan. Information extraction from invoices using graph neural networks. *B.Sc. Thesis, KTH Royal Institute of Technology*, pp. 1–80, 2023.

[26] J. Guo and X. Wang. A structured recognition method for invoices based on StrucTexT model. *Proc. Int. Conf. on Applied Automation in Document Processing*, pp. 123–134, 2023.

[27] P. Damodaran, P. Singh, and J. Achankuju. Zero-shot task transfer for invoice extraction via class-aware QA ensemble. *Proc. Int. Conf. on Document Analysis and Recognition*, pp. 1–12, 2021.

212-15-4196

ORIGINALITY REPORT

7%	5%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Submitted to United International University Student Paper	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	arxiv.org Internet Source	<1%
5	Submitted to Associatie K.U.Leuven Student Paper	<1%
6	Andrei-Stefan Bulzan, Cosmin Cernazanu-Glavan. "Object Detection in Invoices", 2022 26th International Conference on System Theory, Control and Computing (ICSTCC), 2022 Publication	<1%
7	"Document Analysis and Recognition - ICDAR 2024", Springer Science and Business Media LLC, 2024 Publication	<1%
8	www.frontiersin.org Internet Source	<1%
9	www.mdpi.com Internet Source	<1%
10	Submitted to City University of Hong Kong Student Paper	<1%

20% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Detection Groups

47 AI-generated only 20%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.