LLM-centric pipeline for information extraction from invoices

Faiza Loukil, Sarah Cadereau, Hervé Verjus, Mattéo Galfré, Kavé Salamatian, David Telisson, Quentin Kembellec, Olivier Le Van

LLM-centric pipeline for information extraction from invoices

Faiza Loukil*, Sarah Cadereau*†, Hervé Verjus*, Mattéo Galfre†, Kavé Salamatian*,
David Telisson*, Quentin Kembellec†, Olivier Le Van†
*LISTIC, University Savoie Mont Blanc
Annecy-le-Vieux, France
{firstname.lastname}@univ-smb.fr
†Cegedim, Lyon, France
{firstname.lastname}@cegedim.com

***Abstract*—Extracting information from digital documents is an evolving area of research, especially with the recent advances in artificial intelligence and computer vision. Recently, Large Language Models (LLMs) have shown remarkable performance in various natural language processing tasks, including data extraction from documents. However, the accuracy of these models can be significantly affected when dealing with large or complicated documents due to the inherent complexity and variability of rich formats. In this paper, we target a specific type of complex document: financial invoices.**

OCR technology extracts editable and searchable data from different types of documents transformed into an image, *e.g.*, scanned documents, and PDFs. However, OCR is highly sensitive to noise and image mis-alignment that frequently results into wrong extraction of texts. Moreover, OCR cannot understand the structure of a document, and leverage it to understand the semantic of the document's content to extract structured information from document. OCR is therefore considered as a preprocessing step that need to be completed with further processing. In this paper, we use text-based LLMs to enrich the outputs of Optical Character Recognition (OCR) applied to documents to extract structured information from financial invoices. We show here, that by fusing OCR engines, including Tesseract and DocTR, with the two open-source LLM models, Llama3 and Mistral, we significantly improve the accuracy and reliability of information extraction operations on two datasets featuring business documents: SROIE and FATURA datasets.

***Index Terms*—Invoice Processing, Large Language Models, Information Extraction.**

I. Introduction

Today's business environment faces organizations with an ever-increasing volume of documents that need to be processed. Thus, automatizing the capture of structured information from business documents, not only saves a considerable amount of time by minimizing human intervention, but can create business value. For instance, extracting in structured format the information embedded in invoices can be used for business analysis, for example monitoring purchasing behavior [1].

Extracting information from invoices is therefore a major target of applicative research. Recent years have witnessed remarkable progress in artificial intelligence and computer vision, leading to advances in automated document understanding. This has been applied to invoices' information extraction following three main approaches: (i) graph-based approaches [2] [3], which are computationally intensive, particularly for large or complex invoices, and capture spatial relationships between different text elements on an invoice, (ii) deep learning-based approaches [4] [5] [6], which use neural networks calibrated over large amount of labeled training data with high-quality input images, to learn relevant features from invoice images using neural networks, and more recently (iii) large language models (LLMs)-based approaches [7] [8] [9], which take text resulting from OCR, for mono-modal case, or directly the scanned image for multi-modal case, as input and generate structured one-dimensional text sequence. In this paper, we focus on the mono-modal case that accepts text resulting from OCR and output text sequences. The main reason is that we are targeting in this work a practical and affordable system that can be easily deployed, while multi-modal systems are larger and more complex.

This study proposes a two-stage LLM-centric pipeline for structured information extraction from invoices. In the first stage, we extract invoices' content with a relatively simple Optical Character Recognition (OCR) engine, such as Tesseract [10] or DocTR [11], which are are well-known open source OCR frameworks that can use non-commercial engine, to extract from images containing text the embedded text. In the second stage, the text sequence resulting from the OCR is combined with knowledge coming from invoice processing to create, in several steps, prompts that are presented to a pre-trained LLM that solves the invoice comprehension task without further fine-tuning. We validate the proposed pipeline over two datasets containing invoices, namely SROIE [12] and FATURA datasets [13], and using different models supported by Ollama [14], including Llama3 and Mistral.

The rest of this paper is organized as follows. Section II introduces some key concepts relevant to information extraction from business invoices. Section III presents the studies about information extraction from invoices that are relevant to our proposed approach. Section IV describes the proposed LLM-centric pipeline for extracting invoice information. Section V discusses the implementation details and the proposal's performance. Finally, Section VI concludes the paper with future research directions.

II. Background

A computer system automating invoice processing needs to incorporate a wide range of software tools to handle all the tasks. We provide a brief overview of the two key components of such a system : (i) Optical Character Recognition (OCR), which is used to extract data from scanned or PDF invoices, making them searchable and understandable, and (ii) Large Language Models (LLMs), that enable better document understanding and information retrieval, in particular the identification of specific pieces of information within a document, such as vendor information, invoice numbers, product tables, *etc.*

A. OPTICAL CHARACTER RECOGNITION (OCR)

Optical Character Recognition (OCR) is a set of technologies used to automatically identify and extract text from images like scanned documents, screenshots, or formatted documents like PDFs. The evolution of OCR technology spans over a century and continues to evolve, with several notable trends emerging from recent research as multilingual recognition and different font styles, as well as handwritten characters written with any instrument. However, OCR technology still faces several challenges when it comes to matching human reading abilities [15]. Current OCR systems include artificial intelligence and deep learning to improve OCR's ability to recognize text in complex layouts and environments and learn from mistakes and improve over time. Tesseract [10] that is an open-source OCR engine known for its accuracy, and versatility, as well as DocTR [11], a deep learning-based OCR system designed to extract efficiently text from complex document layouts, are examples of modern OCR systems that integrate artificial intelligence to improve OCR capabilities and performance.

B. Large Language Models (LLMs)

Large Language Models (LLMs) are advanced computational models built over deep transformers architecture. They use massive datasets to achieve general-purpose language understanding and generation [16]. The evolution of LLMs can be traced through several key developments, such as the introduction of the transformer architecture and its use in BERT (Bidirectional Encoder Representations from Transformers) by Google, which significantly improved performance on various NLP tasks [17], and attention mechanism [18]. These contributions revolutionized natural language processing and prepared the apparition of ChatGPT-2 and GPT-3 released by OpenAI which demonstrated impressive text generation capabilities. In addition to commercial models, a variety of openly available models such as Llama3 or Mistral have been released [14] and are widely used in research. Recently, the LLM paradigm has become dominant in areas using neural networks beyond NLP, as LLM can also serve as generalist agents for ad-hoc problem-solving, without fine-tuning to specific tasks. LLM technology has started to have far-reaching impacts across various industries and applications, including automating invoice processing.

III. Related Work

Several solutions have been proposed in the literature to address the challenge of extracting information from invoices. We summarize the published research into three categories: graph-based approaches [2] [3], deep learning-based approaches [4] [5] [6], and large language models (LLMs)-based approaches [7] [8] [9].

Krieger et al. [2] proposed a graph-based approach for extracting information from invoices by transforming them into a graph structure, where text elements become nodes in the graph. The authors combined machine learning using Graph Neural Networks (GNN) with heuristic rules to generate the final structured text results. Thereafter, the graph embedded in the GNN is used over a dataset of invoices from multiple vendors. Nonetheless, the approach proposed fails to deal with invoices with completely new layouts or formats that are not represented in the training data. Lohani et al. [3] introduced a model-free Graph Convolutional Network (GCN) approach for reading and processing information in a scanned invoice image. The proposed system extracts table information and can extract up to 27 entities of interest without any template information or configuration with good performance. Although the GCN approach is robust, it encounters scalability challenges related to the computational complexity of graph operations when processing large volumes of invoices with complex structures.

Arslan et al. [4] introduced a deep learning-based method based on Convolutional Neural Networks (CNNs) for invoice segmentation, for identifying and extracting fields from invoice images. These fields are used to develop an end-to-end solution doing automated labeling and enabling synthetic invoice generation. They evaluated the proposed system over 1022 manually labeled invoice images. The fine-tuned model outperformed the baseline models with an accuracy of 8.4%. However, the system performance was susceptible to invoice image quality. Yao et al. [5] proposed an invoice information recognition system for Chinese characters by using deep learning techniques. The system combines YOLOv3 (You Only Look Once version 3) for object recognition and CRNN (Convolutional Recurrent Neural Network) for sequence recognition. Although the system rapidly processed invoice content, it still suffered from ineffective recognition in the case of low-resolution and noisy invoices. Hamdi et al. [6] combine (a) textual features extraction, *i.e.* all the words that can be used to define the context of the word to be labeled with (b) layout features, *i.e.*, the position of the word in the document, block and line as well as its coordinates, (c) pattern features, *i.e.*, each input word is represented by a normalized pattern built by substituting each character with a normalized one, and (d) logical features *i.e.*, boolean values, for example, indicating whether the word is a title. They fine-tuned a BERT model to extract relevant information from invoices. Their experiments encompassing different categories of documents showed some limits over invoices, *i.e.*, some invoices' fields are not detected and prediction results are not satisfactory.

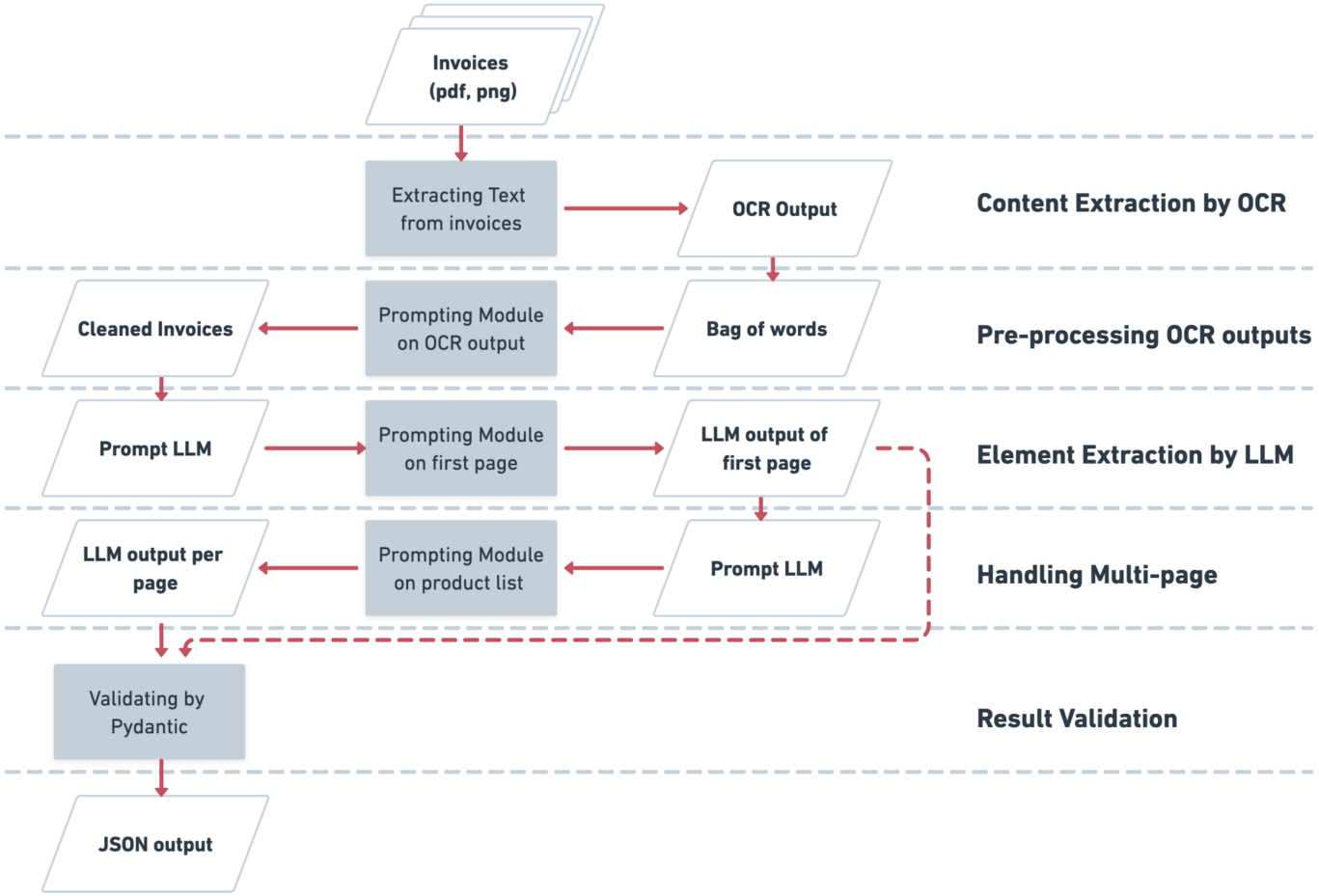


Fig. 1. Overview of the proposed pipeline for extracting invoice information.

The emergence of large language models (LLMs) has led to a major development in invoice information extraction recently. For instance, Perot et al. [7] developed LMDX, a five-step process to extract key-value pairs from documents. This pipeline included OCR, chunking the OCR output, creating prompts, LLM inference, and parsing and decoding of the outputs. Their solution has achieved state-of-the-art performance on two of the most commonly used public benchmarks, VRDU and CORD while remaining data efficient. However, the usage of large language models always entails significant computing costs. Ye et al. [8] proposed a modularized multimodal large language model (MLLM) based on the mPLUG-Owl architecture. The authors first constructed an instruction-tuning dataset featuring a wide range of visual-text understanding tasks. Then, they strengthen the OCR-free document understanding ability by jointly training the model on language-only, general vision-and-language, and document instruction-tuning datasets with a unified instruction-tuning strategy. Despite its better performance, the proposed systems can still generate inaccurate information. Recently, Wang et al. [9] proposed a multimodal approach combining OCR with LLM; they target rich visual documents characterized by their heterogeneous content, irregular layouts, and disjointed text segments. Their pipeline architecture permits the detection of visual information by leveraging the spatial positions and dimensions of text tokens obtained using OCR. They treat the spatial information about the text tokens as a distinct modality, use separate vectors to represent these different modalities, and extend the transformer architecture's self-attention mechanism to compute their disentangled inter-dependencies. Their experiments do not include invoices; instead, they use large amounts of banking visual documents and legal proceedings against the tobacco industry. For our pipeline, we extended and adapted their architecture to invoice-specific aspects.

The above studies show that using OCR and LLM improves the quality of information extracted from business documents. In this work, we combine both approaches and achieve computational saving in LLM while improving OCR performance.

IV. System Design: LLM-centric pipeline for information extraction from invoices

Figure 1 depicts an overview of the system developed in this paper. It begins the processing by using an OCR solution to extract text from a given invoice. This text is thereafter inserted into a prompt template together with a task-specific directive to exclude irrelevant pages. These pre-processing

steps are important to clean the invoice content from unnecessary attributes that could hinder the LLM processing. After that, the filtered invoice content is input into a pre-trained Large Language Model through a defined prompt that asks to extract main information, including supplier's data, total amounts, order numbers, etc. The step of handling multi-pages is used when handling complex structure invoices. The proposed system can extract from the table of products on each page and using the keywords present in the invoice header, the list of products and relevant tax values present in the invoice. This extracted information is then re-inserted into a prompt template to extract global information related to products and amounts. We validate the obtained structured information using Pydantic, a data validation library, that leverages data models describing precisely the expected information, and their format, or generating these models through examples. This last stage standardizes the LLM output as much as possible. Finally, the prepared prompt is then fed into the LLM and the answer is encoded in JSON format. Our pipeline architecture implementation adapts and extends the software framework developed in [19] with specific enhancements, *i.e.* removing irrelevant extra pages.

V. Evaluation and Results

In this section, we evaluate the information extraction performance of the proposed LLM-centric pipeline using two real-world datasets. We present the experimental details and corresponding discussions, including dataset description, the used evaluation metrics, and a detailed comparison of Llama3, Llama3-8B, and Mistral-7B models with respect to information extraction results.

A. Data Description

We evaluate our approach using both the SROIE [12] and the FATURA [13] datasets. The SROIE (Scanned Receipts OCR and Information Extraction) dataset [12] provides a large dataset used to evaluate the robustness of relevant information extraction methods from scanned receipts. It consists of 1000 scanned receipts along with their corresponding OCR results. The objective of using this dataset is to extract four main information for each scanned receipt, namely company name and address, date, and the total. The second used dataset is the FATURA dataset [13], which is a good resource for document analysis, primarily focused on invoice documents. It has 10,000 invoice images in 50 various layouts, making it the largest publicly available image library of invoices to date. FATURA offers annotated images of invoice papers with different layouts. The basis for this choice is twofold. On the one hand, this dataset is intended to address the problems of document-related tasks that require both textual information and exact bounding box annotations to distinguish between document items. On the other hand, it is free and includes detailed benchmarks for various document analysis scenarios.

B. Evaluation Metric

To evaluate our approach's performance in extracting information from invoices, we calculate the following metrics:

- *Task Completion Time*: measures the time (in seconds) taken to complete one specific task-request. We separate the OCR time and LLM time.
- *Text-extraction Accuracy*: measures the accuracy of text extracted from the invoices using LLM.
- *Precision*: measures how many positive predictions are correct (true positives).
- *Recall*: measures how many positive cases the classifier correctly predicted, over all the positive instances in the data.
- *F1-Score*: measures both ratios (precision and recall) in a balanced way, requiring both to have a higher value for the F1-score value to rise.

Three string similarity metrics are used in this study. Two of them are non-semantic: Jaro-Winkler, which is based on character matching and transpositions, and Sorensen-Dice, which is based on n-grams[1]. The third method is a semantic one called all-MiniLM-L12-v2 library[2], that is specifically designed for semantic similarity through a medium-sized model.

C. Experimental Results

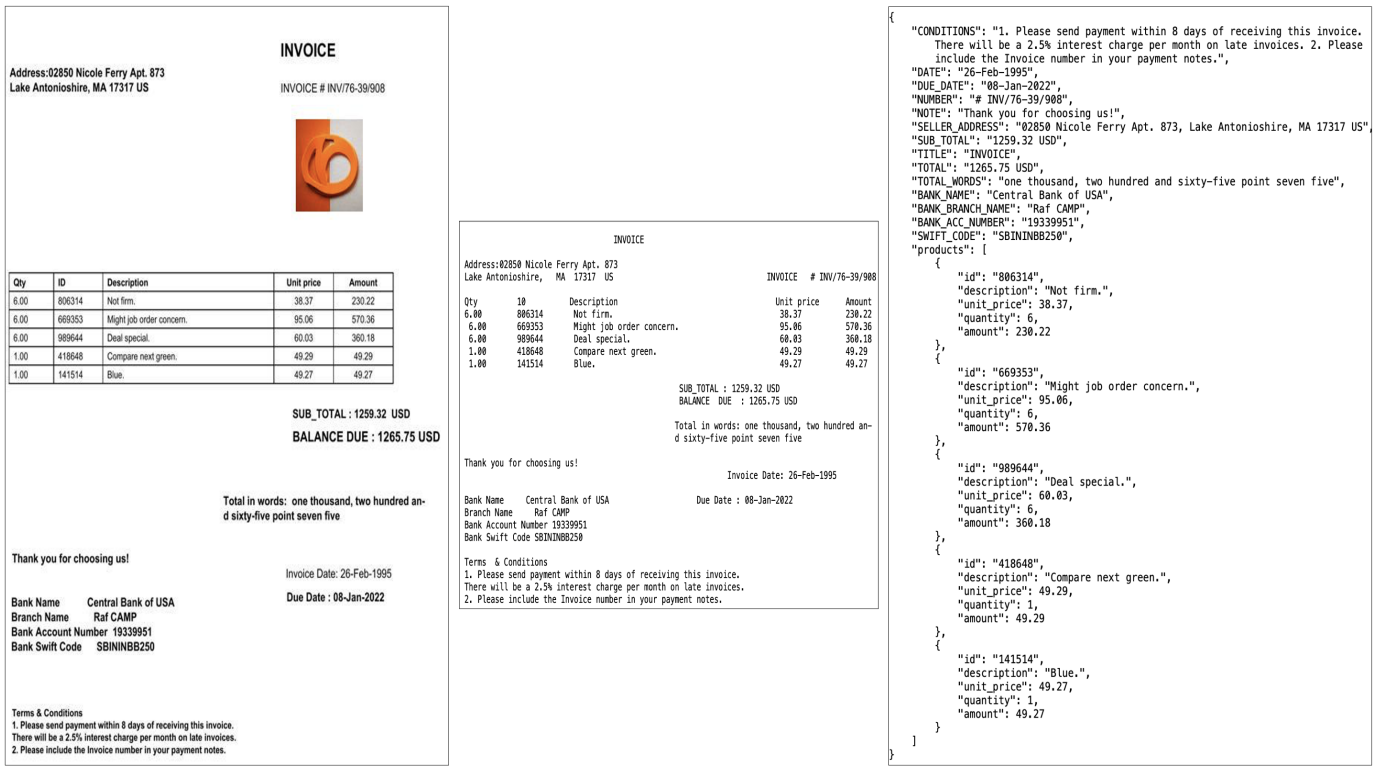
We conduct all experiments using Ollama API [14], which is an open-source platform that allows large language models (LLMs) to run locally on any machine. It provides a simple API for creating, running, and managing models, as well as a library of pre-built models that can be easily used in a variety of applications. In our case, we pulled and used locally the standard Llama3 model, the Llama3 version with 8B parameters, and the high-performance Mistral language model with 7B parameters. We also use for the OCR stage the Tesseract [10] framework, an open-source optical character recognition (OCR) engine that allows images containing text to be converted to machine-encoded text. We have also used DocTR (Document Text Recognition) [11] that uses deep learning for OCR, and is claimed to enable advanced document understanding, but has a commercial license.

Given the OCR results, we enrich the LLM prompts to improve invoice understanding performance. We perform all experiments on a machine with an Intel Core i9-13900 CPU and NVIDIA RTX 6000 Ada Generation with 48 GO GDDR6 memory.

In the rest of this section, we report the results on both datasets with the chosen OCR engines and LLM models. The results in Figure 2 show how our system generates a textual document representation from a random invoice sample from the FATURA dataset. On the left side, the original scanned invoice appears as an image or a pdf. In the middle, the OCR output is obtained after transforming the invoice image into a machine-readable text, though it might have some errors or misalignment. The LLM output is shown on the right side. It improves the OCR text by fixing errors and structuring the data into JSON format as required in the executed prompt.

[1]Available in the Python library *textdistance* https://github.com/life4/textdistance

[2]Implemented in the python library *sentence-transformers* https://www.sbert.net/docs/sentence_transformer/pretrained_models.html



INVOICE

Address:02850 Nicole Ferry Apt. 873
Lake Antonioshire, MA 17317 US

INVOICE # INV/76-39/908

Qty	ID	Description	Unit price	Amount
6.00	806314	Not firm.	38.37	230.22
6.00	669353	Might job order concern.	95.06	570.36
6.00	989644	Deal special.	60.03	360.18
1.00	418648	Compare next green.	49.29	49.29
1.00	141514	Blue.	49.27	49.27

SUB_TOTAL : 1259.32 USD

BALANCE DUE : 1265.75 USD

Total in words: one thousand, two hundred and sixty-five point seven five

Thank you for choosing us!

Invoice Date: 26-Feb-1995

Due Date : 08-Jan-2022

Bank Name Central Bank of USA
Branch Name Raf CAMP
Bank Account Number 19339951
Bank Swift Code SBININBB250

Terms & Conditions
1. Please send payment within 8 days of receiving this invoice.
There will be a 2.5% interest charge per month on late invoices.
2. Please include the Invoice number in your payment notes.

```
INVOICE

Address:02850 Nicole Ferry Apt. 873
Lake Antonioshire,   MA  17317  US                    INVOICE   # INV/76-39/908

Qty         10          Description                    Unit price      Amount
6.00        806314       Not firm.                      38.37          230.22
 6.00       669353       Might job order concern.       95.06          570.36
 6.00       989644       Deal special.                  60.03          360.18
 1.00       418648       Compare next green.            49.29          49.29
 1.00       141514       Blue.                          49.27          49.27

                                        SUB_TOTAL : 1259.32 USD
                                        BALANCE  DUE  : 1265.75 USD

                                        Total in words: one thousand, two hundred an-
                                        d sixty-five point seven five

Thank you for choosing us!

                                                    Invoice Date: 26-Feb-1995

Bank Name     Central Bank of USA                   Due Date : 08-Jan-2022
Branch Name     Raf CAMP
Bank Account Number 19339951
Bank Swift Code SBININBB250

Terms  & Conditions
1. Please send payment within 8 days of receiving this invoice.
There will be a 2.5% interest charge per month on late invoices.
2. Please include the Invoice number in your payment notes.
```


```
{
    "CONDITIONS": "1. Please send payment within 8 days of receiving this invoice.
       There will be a 2.5% interest charge per month on late invoices. 2. Please
       include the Invoice number in your payment notes.",
    "DATE": "26-Feb-1995",
    "DUE_DATE": "08-Jan-2022",
    "NUMBER": "# INV/76-39/908",
    "NOTE": "Thank you for choosing us!",
    "SELLER_ADDRESS": "02850 Nicole Ferry Apt. 873, Lake Antonioshire, MA 17317 US",
    "SUB_TOTAL": "1259.32 USD",
    "TITLE": "INVOICE",
    "TOTAL": "1265.75 USD",
    "TOTAL_WORDS": "one thousand, two hundred and sixty-five point seven five",
    "BANK_NAME": "Central Bank of USA",
    "BANK_BRANCH_NAME": "Raf CAMP",
    "BANK_ACC_NUMBER": "19339951",
    "SWIFT_CODE": "SBININBB250",
    "products": [
        {
            "id": "806314",
            "description": "Not firm.",
            "unit_price": 38.37,
            "quantity": 6,
            "amount": 230.22
        },
        {
            "id": "669353",
            "description": "Might job order concern.",
            "unit_price": 95.06,
            "quantity": 6,
            "amount": 570.36
        },
        {
            "id": "989644",
            "description": "Deal special.",
            "unit_price": 60.03,
            "quantity": 6,
            "amount": 360.18
        },
        {
            "id": "418648",
            "description": "Compare next green.",
            "unit_price": 49.29,
            "quantity": 1,
            "amount": 49.29
        },
        {
            "id": "141514",
            "description": "Blue.",
            "unit_price": 49.27,
            "quantity": 1,
            "amount": 49.27
        }
    ]
}
```


Fig. 2. Element extraction by Llama3 on a random invoice sample from the FATURA dataset: original scanned invoice (left), OCR output (middle), and LLM output (right).

We first compare the execution time performance of the two alternative OCR engines, namely Tesseract [10] and DocTR [11]. The results in Table I show that overall each dataset can be processed in a few minutes and that the average OCR task completion time per invoice (in seconds) is acceptable. Variations in task completion time across different datasets and different invoices are mainly explained by the difference in complexity of the given invoice structures and the disparity of formats.

TABLE I
COMPARISON OF TASK COMPLETION TIME TAKEN BY TESSERACT AND DOCTR TO EXTRACT INVOICE TEXT FROM SROIE AND FATURA DATASETS.

		OCR TIME (sec)	
		Total	AVG
SROIE (973 invoices)	**Tesseract**	602,09	0,62
	DocTR	1271,52	1,31
FATURA (1000 invoices)	**Tesseract**	483,51	0,48
	DocTR	1141,14	1,14

Table II and Table III show the comparison of performance metrics over the SROIE and FATURA datasets, of the applications of the three LLM models, namely Llama3 (4,7 GB), llama3-8b-instruct-q5_K_M (5,7 GB), and mistral-7b-instruct-v0,3-q5_K_M (5,1 GB) when they are fed by the output of the two alternative OCR engines, Tesseract [10] and DocTR [11]. For both datasets, we show that Llama3 performs slightly faster than Llama3-8B and Mistral-7B, *i.e.*, the average Llama3 processing time for the SROIE dataset is 0.98 seconds, compared to 1.04 seconds for Llama3-8B and 1.20 seconds for Mistral-7B. Furthermore, Tesseract and DocTR display comparable speed performance across models with little changes. Therefore, the commercial DocTR does not achieve higher speed and throughput than an open-source OCR engine, like Tesseract.

Table II shows that accuracy achieved by the combination of the Mistral-7B LLM model and the DocTR OCR engine is the best over the SROIE dataset. While in Table III, this is the combination of Tesseract and Llama3 that achieves the best accuracy over the FATURA dataset. In both cases, the type of the LLM model used is not very important with slight variation among them. This means that the use of a smaller model like Llama3 will achieve most of the gain of the LLM at a lower implementation cost and hardware cost. The main difference is with the choice of the OCR system. This can be related to the fact that the SROIE is a dataset used for robustness analysis and concentrates on harder cases, than the FATURA dataset, which is larger and contains a more typical invoice. Overall, we can conclude that using the combination of Tesseract and Llama3 is combining for large datasets the lowest execution time, and an acceptable performance (with little loss compared to the more complex LLMs). However, it may be necessary to manually analyze the cases in the SROIE dataset that have poor performance to understand how to improve the performance over them.

TABLE II
EVALUATION RESULTS FOR SROIE DATASET AND COMPARISON WITH LLAMA3 (4,7 GB), LLAMA3-8B (5,7 GB), AND MISTRAL-7B (5,1 GB) MODELS.

			Llama3		**Llama3-8B**		**Mistral-7B**	
			Tesseract	DocTR	Tesseract	DocTR	Tesseract	DocTR
	LLM TIME (sec)	Total	**951,07**	952,94	1010,64	1015,95	1167,23	1111,25
		AVG	**0,98**	0,98	1,04	1,04	1,20	1,14
SIMILARITY METRICS	all-MiniLM-L12-v2	Accuracy	0,66	0,82	0,66	0,83	0,68	**0,85**
		Precision	0,67	0,83	0,67	0,83	0,68	0,85
		Recall	0,92	0,99	0,92	1,00	0,94	1,00
		F1-Score	0,75	0,89	0,75	0,89	0,77	0,91
	Jaro-Winkler	Accuracy	0,71	0,86	0,71	0,87	0,73	**0,89**
		Precision	0,72	0,87	0,72	0,88	0,73	0,90
		Recall	0,93	1,00	0,93	1,00	0,94	1,00
		F1-Score	0,79	0,92	0,79	0,92	0,80	0,94
	Sorensen-Dice	Accuracy	0,65	0,83	0,64	0,81	0,68	**0,86**
		Precision	0,67	0,84	0,66	0,82	0,70	0,87
		Recall	0,92	1,00	0,92	1,00	0,93	1,00
		F1-Score	0,75	0,89	0,74	0,88	0,77	0,92

TABLE III
EVALUATION RESULTS FOR FATURA DATASET AND COMPARISON WITH LLAMA3 (4,7 GB), LLAMA3-8B (5,7 GB), AND MISTRAL-7B (5,1 GB) MODELS.

			Llama3		**Llama3-8B**		**Mistral-7B**	
			Tesseract	DocTR	Tesseract	DocTR	Tesseract	DocTR
	LLM TIME (sec)	Total	**2183,39**	2187,40	2337,83	2313,18	3363,46	3429,99
		AVG	**2,18**	2,19	2,34	2,31	3,36	3,43
SIMILARITY METRICS	all-MiniLM-L12-v2	Accuracy	**0,90**	0,75	**0,90**	0,75	0,86	0,70
		Precision	0,93	0,78	0,93	0,78	0,90	0,74
		Recall	0,96	0,94	0,96	0,94	0,95	0,93
		F1-Score	0,94	0,85	0,94	0,85	0,92	0,82
	Jaro-Winkler	Accuracy	**0,92**	0,84	**0,92**	0,84	0,89	0,81
		Precision	0,95	0,89	0,95	0,89	0,93	0,85
		Recall	0,96	0,94	0,96	0,95	0,95	0,94
		F1-Score	0,95	0,91	0,96	0,91	0,94	0,89
	Sorensen-Dice	Accuracy	**0,91**	0,83	**0,91**	0,83	0,88	0,79
		Precision	0,95	0,87	0,95	0,87	0,93	0,83
		Recall	0,96	0,94	0,96	0,95	0,95	0,93
		F1-Score	0,95	0,90	0,95	0,90	0,93	0,87

VI. CONCLUSION

Large Language Models (LLMs) leverage vast amounts of text data to understand context and semantics, making them highly effective at extracting information from digital invoices. In this paper, we enrich LLM prompts with OCR information to improve invoice understanding performance. The proposed pipeline requires only preprocessing of the invoice text and prompts without additional fine-tuning. Experimental results show that the open-source LLM models used, Llama3, Llama3-8B, and Mistral-7B, perform well on both SROIE and FATURA datasets when enriched with output from OCR engines, such as Tesseract or DocTR. The synergy between OCR and LLMs allows for better handling of complex invoice structures and varying formats, ensuring more accurate extraction of essential information from invoices. For future research, we intend to add layout information to LLM prompts for instruction-tuned LLMs to improve invoice understanding performance. We also plan to investigate the benefits and trade-offs of incorporating multi-modal large language models for extracting information from invoices.

APPENDIX
PROMPT TEMPLATE

This listing gives an overview of the used prompt template.

```
You receive invoice content.

You must extract specific information
    from the invoice and return them in a
    well-structured JSON format FRAMED
    WITH ```.
The information to be extracted concerns:
$$$
<<<Extracted Content by an OCR>>>
$$$

Here are the JSON fields required,
    described with comments following --,
    but DO NOT INCLUDE COMMENTS in your
    generated JSON:
$$$
<<<Generated JSON Schema specific to the
    Extracted Content>>>
$$$
```

ACKNOWLEDGMENT

This research was supported by funding from the Cegedim company, which provided us with computing resources.

REFERENCES

[1] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "Rfm-based repurchase behavior for customer classification and segmentation," *Journal of Retailing and Consumer Services*, vol. 61, p. 102566, 2021.

[2] F. Krieger, P. Drews, B. Funk, and T. Wobbe, "Information extraction from invoices: a graph neural network approach for datasets with high layout variety," in *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*. Springer, 2021, pp. 5–20.

[3] D. Lohani, A. Belaïd, and Y. Belaïd, "An invoice reading system using a graph convolutional network," in *Asian Conference on Computer Vision*. Springer, 2019, pp. 144–158.

[4] H. Arslan, Y. E. Işık, and Y. Görmez, "A deep learning-based solution for digitization of invoice images with automatic invoice generation and labelling," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 27, no. 1, pp. 97–109, 2024.

[5] X. Yao, H. Sun, S. Li, and W. Lu, "Invoice detection and recognition system based on deep learning," *Security and Communication Networks*, vol. 2022, no. 1, p. 8032726, 2022.

[6] A. Hamdi, E. Carel, A. Joseph, M. Coustaty, and A. Doucet, "Information Extraction from Invoices," in *International Conference on Document Analysis and Recognition ICDAR 2021*, ser. Lecture Notes in Computer Science, vol. 12822. Lausanne, Switzerland: Springer International Publishing, Sep. 2021, pp. 699–714. [Online]. Available: https://hal.science/hal-03418385

[7] V. Perot, K. Kang, F. Luisier, G. Su, X. Sun, R. S. Boppana, Z. Wang, J. Mu, H. Zhang, and N. Hua, "Lmdx: Language model-based document information extraction and localization," *arXiv preprint arXiv:2309.10952*, 2023.

[8] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, Y. Dan, C. Zhao, G. Xu, C. Li, J. Tian *et al.*, "mplug-docowl: Modularized multimodal large language model for document understanding," *arXiv preprint arXiv:2307.02499*, 2023.

[9] D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, and X. Liu, "DocLLM: A layout-aware generative language model for multimodal document understanding," Dec. 2023, arXiv:2401.00908 [cs]. [Online]. Available: http://arxiv.org/abs/2401.00908

[10] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[11] H. Liao, A. RoyChowdhury, W. Li, A. Bansal, Y. Zhang, Z. Tu, R. K. Satzoda, R. Manmatha, and V. Mahadevan, "Doctr: Document transformer for structured information extraction in documents," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 584–19 594.

[12] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "Icdar2019 competition on scanned receipt ocr and information extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1516–1520.

[13] M. Limam, M. Dhiaf, and Y. Kessentini, "Fatura: A multi-layout invoice image dataset for document analysis and understanding," *arXiv preprint arXiv:2311.11856*, 2023.

[14] Ollama, "Ollama," 2023. [Online]. Available: https://ollama.com/

[15] T. Saout, F. Lardeux, and F. Saubion, "An overview of data extraction from invoices," *IEEE Access*, 2024.

[16] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[19] Julio/enoch3712, "Extractthinker," July 2024. [Online]. Available: https://github.com/enoch3712/ExtractThinker