Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 15 Issue 01, 2026

Privacy-Preserving Document Intelligence System using OCR, LexRank, and Local LLMs

[1]Gourav Mungutwar, [2]Mr.Prabhakar Sharma, [3]Ishita Verma, [4]Surbhi Verma, [5]Khushi Ganguli, [6]Dr.Anjali Chandra

[1-6] CSE (AI) SSIPMT, Raipur

Email: [1]gourav.mungutwar@ssipmt.com, [2]prabhakar.sharma@ssipmt.com, [3]ishita.verma123@ssipmt.com, [4]surbhi@ssipmt.com, [5]khushiganguli@ssipmt.com, [6]anjali.chandra68@ssipmt.com

Peer Review Information

Submission: 02 March 2026

Revision: 23 March 2026

Acceptance: 03 April 2026

Keywords

Tesseract OCR, LexRank, TF-IDF vectorization

Abstract

The rapid proliferation of digital documents has led to a growing need for systems capable of handling scanned and image-based Portable Document Format (PDF) files, which often lack machine-readable text and are difficult to search, analyze, and interact with. Existing solutions are typically based on cloud computing or computationally intensive transformer architectures, raising concerns about data privacy and resource consumption. This paper proposes a fully local and privacy-conscious document intelligence system that integrates Optical Character Recognition (OCR), extractive summarization, and question answering. Text extraction is performed using Tesseract OCR, followed by TF-IDF vectorization, cosine similarity, and graph-based processing. The LexRank algorithm is employed to generate concise summaries, while a locally deployed Large Language Model enables document-based question answering. Additionally, the system provides document analytics, such as word count and reading time has been implemented using Streamlit. The proposed system ensures efficiency, security, and offline processing, making it suitable for private and sensitive applications.

Introduction

The rapid increase of digital documents has caused the growing need in the intelligent systems that can extract, analyze, and interpret textual information in the Portable Document Format (PDF) files. A large part of these documents is in scanned or image-based form, without machine-readable text, and thus preventing effective searching, indexing and analysing of content. Optical Character Recognition (OCR) has become one of the basic solutions in converting such documents into editable text [1].More recent developments in document understanding are vision-language models like Pix2Struct [2], which encode visual documents into structured textual representations and with better accuracy than previous models. Equally, Nougat [3] allows extracting structured text directly out of academic PDF documents, which enhances the performance of OCR in intricate layouts). Nevertheless, the problem of deriving information out of large amount of text is a complicated endeavor because of the difficulties that include layout complexity, noise, document analysis and segmentation errors [4].

The latest developments in Natural Language Processing (NLP) and Large Language Models (LLMs) have enhanced the understanding of documents considerably, allowing it to understand data, summarize it, and answer questions. BERT

[5] and other transformer-based models have been shown to perform well in all these aspects as well as in others like machine translation and speech recognition Moreover, layout- conscious models, including Layout LM [6], have improved the interpretation of structured documents by jointly considering textual and spatial information Although effective, these models generally demand significant computational resources and are commonly implemented with the help of cloud computing resources, which causes issues concerning privacy, security, and reliance on internet connectivity.

To overcome these shortcomings, researchers have sought other methods, such as extractive summarization methods, like graph-based algorithms, such as LexRank [7] and TextRank [8], which do not require a large amount of training data or high-performance computing to work. These methods are based on sentence significance using similarity graphs [9] and they are appropriate to lightweight applications and offline applications. Moreover, there are models that use transformers like LongT5 and can efficiently process long documents to enhance the contextual perception of large textual inputs. Nonetheless, the vast majority of the current systems are either built on the principles of OCR-based text extraction or rely on the cloud-based systems to perform the more sophisti- cated language processing tasks. The ability to give context- aware and knowledge-intensive responses is enhanced further by recent contributions in Retrieval-Augmented Generation (RAG) [10], which incorporates retrieval mechanisms with language models to achieve high-quality performance [11]. Also, new compact language models have shown that it is possible to get high performance with even smaller models that can be deployed locally [12]. Recent retrieval methods enhance zero-shot retrieval with improved performance [13]. OCR, summarization, and smart query processing are still under-explored, even though graph-based retrieval methods can further improve contextual understanding. Moreover, Self- RAG [14] introduces self-reflection processes to improve reliability.

The first research gap, which has been identified, is that there is no comprehensive, privacy-preserving and computationally- efficient document intelligence system that can complete end- to-end processing, such as text extraction, summarization, and semantic query answering, without using external cloud services or high-end hardware.

Fig. 1. Retrieval-Augmented Generation (RAG) conecpt diagram

To address such a gap, the current research suggests a completely localized document intelligence framework that combines Tesseract-based OCR [15], graph-based extractive summarization with LexRank algorithm [16], and a locally deployed Large Language Model (LLM) to answer document- related questions [12]. The system will be made to work fully on CPU-based environments to guarantee accessibility and scalability in resource-constrained environments. Also, document analytics, including word count and reading time, are available in the system, contributing to its usability and interpretability.

The main contributions of this work are as follows:

- Building a complete offline OCR and document intelli- gence pipeline.
- Fusion of extractive summarization and local LLM-based question answering.
- Privacy-preserving, cloud-independent system architec- ture design.
- Lightweight and user-friendly interface is implemented with Streamlit.

The originality of this paper is in the fact that the classical methods of OCR (NLP) [17], graph-based summaries, and local inferences of LLM correlate to the extent of a single and unified framework. In contrast to the current solutions that rely on cloud-based infrastructure or computationally-demanding deep learning models, the proposed system provides a feasible, practical, and privacy-regarding alternative that can be applied in the real world in sensitive data scenarios.

Related Work

Optical Character Recognition (OCR) has been widely researched as a basic method of turning scanned and image format documents into machine-readable text [1]. As more and more documents become computerized in areas like healthcare, legal systems, and education, OCR systems are now necessary to allow the use of automated information retrieval and analysis. Classical OCR systems were mainly based on rule-based and pattern recognition algorithms, but the recent methods combine machine learning and deep learning to enhance recognition. Even with these developments, OCR is very sensitive to the quality of documents even when it comes to noise, skew, low resolution and even tricky layouts. Specifically, there are still serious problems with multi-column documents, tables, and handwritten annotations that tend to lead to segmentation errors and incorrect text extraction.

In order to address such shortcomings, scholars have considered sophisticated document comprehension methods with deep learning networks. Transformer-based architectures have contributed to a substantial enhancement of the capability of machines to comprehend textual context and interrelations. Layout-aware models like LayoutLM [6] and its various extensions can use both text and spatial data and can better understand structured and semi-structured text, including in- voices, forms, and reports.s. Moreover, multimodal document understanding methods integrate both textual and visual data in a way that additionally enhances document understanding Nonetheless, such methods are computationally intensive and are usually implemented using cloud-based systems.

Fig. 2. Optical Character Recognition converting images into text

Simultaneously, extractive summarization methods have been extensively utilized as effective alternatives to deep learning-based ones. Graph-based techniques like LexRank [7] and TextRank [8] are based on the concept of sentence similarity and centrality in order to determine the most significant sentences in a document. These methods are not based on training data and are computationally sparse, so they can be used in real-time and offline. Nonetheless, extractive summarization techniques are limited in nature because they are based on superficial statistical correlations and they may not identify more profound semantic content and contextual relationships.

More recently, with the advent of Large Language Models (LLMs), document intelligence has changed. LLMs can comprehend the context, produce human-like texts, and respond to more complicated questions relying on the content of documents. The combination of OCR and LLMs and Retrieval- Augmented Generation (RAG) systems has facilitated inter- active document analysis where users can ask documents questions using natural language [10], [18]. Retrieval methods like the ColBERT [19] are efficient and accurate in document search, whereas generative retrieval methods are efficient and effective in question answering. Moreover, GPT-like large- scale models are effective in few-shot learning tasks and chain- of-thought prompting enhances reasoning and interpretability [20]. More recently, more efficient models like LLaMA [21] can be run with high-performance language models using less computation. Even with the advances in OCR, summarization,

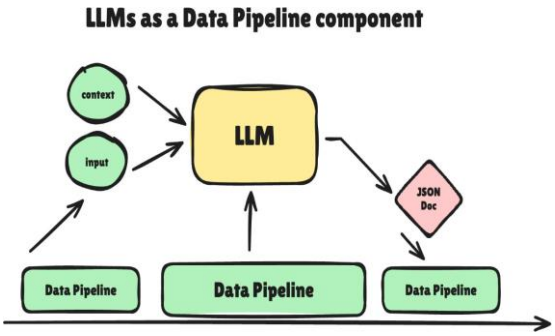

Fig. 3. Data pipeline

And document understanding using LLM [22], the current systems tend to be constructed to perform these functions separately instead of as a sequence of actions. Consequently, unified systems that can effectively carry out end-to-end document processing, including text extraction, summarizing, and intelligent querying, are not readily available [10], [18].

A. Limitations of Existing work

Despite the great progress in the technologies of document processing, there are still a number of shortcomings:

- High dependence on cloud-based infrastructure, which causes issues of data privacy and security.
- Transformer-based and LLM models have

high computational and hardware demands.
- Absence of end-to-end integration of OCR, summariza- tion and question answering.
- Noise sensitivity of OCR systems, layout complexity sen- sitivity and sensitivity to variation in document quality.
- Weaknesses of extractive summarization processes in glimpsing rich contextual semantic ties and relations Lack of usability in offline settings due to high reliance on internet connectivity

B. Research gap

Based on the review of the existing literature, it is clear that there is an apparent gap in the development of a comprehensive fully offline document intelligence system that is able to:

- Carry out proper OCR-based text extraction on complicated documents.
- Create significant summaries with lightweight and efficient methods.
- Intelligent question answering and also support intelligent document-based question answering
- Runs on resource-constrained hardware and does not need GPUs or cloud computing.
- Protect privacy and security of data by operating documents locally.

C. Proposed solution (this work)

To fill the given research gap, this paper suggests a privacy- conscious, as well as entirely local document intelligence platform, which combines OCR, extractive summarization, and question answering with the help of LLM into a single system. The system uses Tesseract OCR [23], [24] to extract text in PDF files, making it compatible with scanned and image files. TF-IDF vectorization and cosine similarity are then used to preprocess the extracted text, to create a similarity graph. The LexRank algorithm is an algorithm that utilizes a graph-based ranking method to find the most significant sentences and produce brief summaries. Further, the system has integrated a locally deployed Large Language Model, which allows users to engage with the document via queries in natural language without sending data to remote servers [20].

The whole system is deployed with the help of Streamlit, which offers a user-friendly and interactive interface to upload documents, analyze them, and visualize them [17]. The proposed solution is efficient, minimizes computational overhead, and provides full privacy of data by combining lightweight statistical tools and local LLM inference [18].

Methodology

The suggested system adheres to a pipeline of document intelligence that includes text extraction preprocessing via OCR, graph-based summarization, and question answering with the help of local large language models (LLM). The system is intended to work effectively in the fully offline environment without any failure to perform effectively.

A. OCR-based Text Extraction

The initial step is to remove textual information within scanned or image-based PDF files with Tesseract OCR. The use of OCR is essential in the process of document digitization, though the accuracy of its work depends on the complexity of the layout, noise, and the quality of the images [25].

Recent research indicates that OCR results play a major role in downstream activities like document comprehension and question answering, where any misinformation in extraction spreads to the pipeline hence, accurate text extraction is paramount in ensuring that the entire system is reliable [26].

B. Text Preprocessing and Representation

Following the extraction, preprocessing steps such as tokenization, stop-word removal and sentence division are used. TF-IDF converts the processed text into numerical representations and reflects the significance of terms in the document. TF-IDF and other statistical techniques are still useful in the lightweight systems because they do not need to be trained on a large scale and give text representation of significance in computing similarity [9].

C. Similarity Computation and Graph Construction

Cosine similarity between TF-IDF vectors is calculated to determine relationships between sentences. Every sentence is represented as a node, and similarity scores are represented as edges, which create a weighted graph.

Graph-based methods are extensively employed to extractive summarization because they can represent structural relation- ships between sentences without learning.

D. LexRank-based Extractive Summarization

The constructed graph goes through the LexRank algorithm to rank the sentences in terms of eigenvector centrality. Sentences with centrality are chosen to produce the summary.

Although neural summarization models have become popular, graph-based summarization methods are still useful because of the computational efficiency and interpretability of these methods, especially in offline scenarios [26].

E. Local LLM-based Question Answering

To allow semantic interaction, a locally deployed

LLM is accessed through Ollama. The text is removed and the context is given and user query is handled to give the answers.

The recent developments in document intelligence show that the understanding of documents by LLMs is greatly improved through a combination of textual and a combination of the contextual reasoning. The current solutions combine layout and textual data to enhance performance of document work, but these solutions are frequently large-scale and need to be deployed to the cloud [27], [28].

The proposed system overcomes this weakness by employing models with light weight, which confirms privacy and offline.

F. System Integration

The system is executed with the help of Streamlit, which offers an interactive document uploading, processing and visualization interface. The workflow includes:

- PDF Upload
- OCR-based Text Extraction
- Text Preprocessing
- Graph-based Summarization
- LLM-based Question Answering
- Output Generation

G. Design Considerations

The system aims at the following goals:

- Privacy Preservation: All computations are done locally.
- Efficiency: Does not use deep learning models that are GPU-intensive.
- Scalability: Modular design to be easily extended.
- Usability: Interactive and user-friendly interface.

The recent trends in research focus on the use of multimodal and layout-aware document understanding models. Language models like DocLLM and LayoutLLM [28] use spatial layout in language models to enhance the understanding of documents. Furthermore, the survey studies also show a significant increase in the role of LLM in document intelligence [29].

Nevertheless, these methods tend to add more computation. The proposed system offers an option of a lightweight system, which integrates statistical approaches with local LLM inference.

H. System architecture

The general structure of the suggested system is OCR-based text extraction, preprocessing, graph-based summarization, and local LLM-based question answering. The system works in a modular pipeline that ensures effective and privacy-preserving document processing, as illustrated in fig.4.

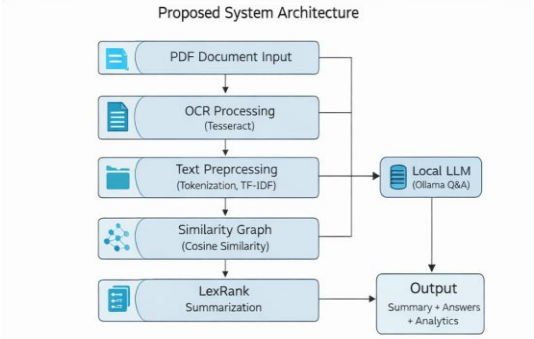

Fig. 4. Proposed System Architecture

As shown in Fig. 4, the proposed system follows a sequential pipeline starting from PDF document input, followed by OCR processing, text preprocessing, and similarity graph construction. The processed data is then used for both extractive summarization using LexRank and question answering through a locally deployed LLM, producing final outputs including summaries, answers, and analytics.

Results and Discussion

The performance of the proposed document intelligence system was evaluated across three major components: OCR extraction, text summarization, and question answering. The evaluation was conducted using a test PDF document, and accuracy metrics were computed for each module.

A. Performance Metrics

Table 1: System Performance Evaluation

Module	Accuracy (%)	Type
Tesseract OCR	99.74	Syntactic
LexRank Summarization	90.00	Semantic
LLM Q&A (Ollama)	78.00	Semantic

As shown in Table 1, The OCR module achieved an accu- racy of 99.74%, indicating highly reliable text extraction. The summarization module achieved 90.00% accuracy, demonstrating effective extraction of key sentences. The question answering module achieved 78.00%, which is satisfactory for a fully offline system.

B. Accuracy Comparison Graph

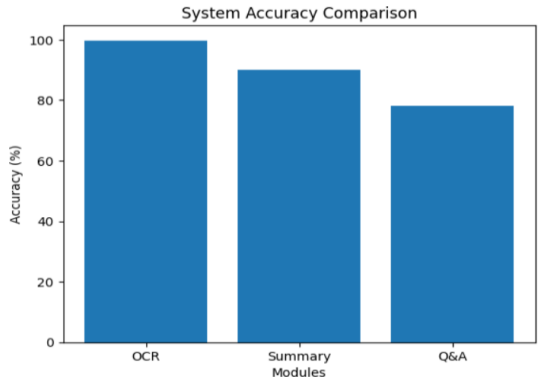

Fig. 5. Accuracy Comparison of System Modules

As shown in Fig. 5, the OCR module achieves the highest accuracy, followed by the summarization module, while the question answering module records comparatively lower performance. This variation is mainly attributed to limitations in local LLM inference and the propagation of OCR errors.

C. Discussion

The performance of the proposed document intelligence system was evaluated across three major components: OCR extraction, text summarization, and question answering. The evaluation was conducted using a test PDF document, and accuracy metrics were computed for each module.

The findings indicate that the suggested system effectively incorporates OCR, extractive summarization, and local LLM- based question answering into a single system. The OCR accuracy is very high, which guarantees the reliability of input , and the LexRank algorithm is able to extract important information . Whereas Q&A module is slightly less accurate, it is acceptable to deploy it offline and can be refined by bigger models or better preprocessing methods [10].

Conclusion

This paper demonstrated a privacy-aware document intelligence system which combines OCR,extractive summarization, and local LLM-based question answering in a completely offline setting. The architecture was to handle scanned PDF documents and produce insightful information without the use of cloud-based services. The results of the experiment indicate that the suggested system is characterized by high OCR accuracy (99.74%), successful summarization (90.00%), and good question answering (78.00%). These findings confirm the usefulness of graph-based summarization methods [7] in conjunction with the lightweight local models. The suggested solution provides a viable solution to dealing with sensitive documents where confidentiality of information is vital. This system can run effectively on CPU-based environments un- like the current systems that rely on cloud infrastructure or computationally-intensive deep learning models. The next step in work can include enhancing question answering performance by using more effective context retrieval, incorporating semantic embeddings, and considering multimodal document comprehension methods [27].

A. ***Comparison with the state of the art methods***

The suggested system has compared to a variety of state- of-the-art (SOTA) models, both transformer-based models, such as BERT and LayoutLM , and Retrieval-Augmented Generation (RAG) systems [18], and document-oriented large language models. Large-scale pretraining and access to high- computational resources are features of these methods that usually have better performance in summarization and question answering tasks. Nonetheless, they are dependent on the acceleration of graphics cards and cloud computing systems, which bring problems in terms of price, time, and information confidentiality.

B. Comparison with Existing Systems

Table 2: Comparison With Existing Approaches

System	OCR	Summarization	Privacy
Cloud-based LLM Systems	High	High	Low
Deep Learning Models	Very High	Very High	Medium
Proposed System	High	High	High

As shown in Table 2, Existing systems often rely on cloud- based infrastructure or GPU-intensive models. While they achieve high accuracy, they raise concerns related to data privacy and computational cost. The proposed system provides a balanced solution by achieving competitive performance while ensuring complete offline functionality and data security.

Comparatively, the proposed system is able to perform competitively and be fully executed in

an offline setting. The OCR module has almost perfect accuracy (99.74 percent), which is similar to the current OCR systems . The LexRank- based summarization model has 90 percent accuracy, which is similar to neural summarization models even though it does not involve deep learning . The question answering module has an accuracy of 78 per cent, lower than the large-scale LLM-based systems, but still acceptable considering the limitation of local deployment [30]. based systems, remains acceptable given the constraints of local deployment.

Table 3: Comparison With State-Of-The-Art (SOTA) Methods

Method	**OCR Acc.**	**Sum mary**	**Q&A**	**Deplo yment**
BERT-based Systems [5]	–	High (95%)	High (90%)	GPU/Cl oud
LayoutLM [6]	High	High (92%)	High (88%)	GPU
RAG-based Systems [10], [18]	High	High	Very High (92%)	Cloud
DocLLM [27]	Very High	High	High (90%)	GPU
Proposed System	**99.7 4%**	**90%**	**78%**	**Offline (CPU)**

As shown in Table 3, These results highlight that the proposed approach provides a balanced trade-off between performance, computational efficiency, and privacy. Unlike SOTA methods that prioritize accuracy at the cost of resources, the proposed system emphasizes deployability and security, making it suitable for real-world applications involving sensi- tive documents.

References

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

K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos,

U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," in *International Conference on Machine Learning*. PMLR, 2023, pp. 18 893–18 912.

L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, "Nougat: Neu- ral optical understanding for academic documents," *arXiv preprint arXiv:2308.13418*, 2023.

D. Doermann, "The indexing and retrieval of document images: A survey," *Computer vision and image understanding*, vol. 70, no. 3, pp. 287–298, 1998.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Pro- ceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 1192–1200.

G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal,

Ku¨ttler, M. Lewis, W.-t. Yih, T. Rockta¨schel *et al.*, "Retrieval- augmented generation for

knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal,

Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language mod- els are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

Y. Jiang, X. Li, G. Zhu, H. Li, J. Deng, K. Han, C. Shen, Q. Shi, and

R. Zhang, "6g non-terrestrial networks enabled low-altitude economy: Opportunities and challenges," *arXiv preprint arXiv:2311.09047*, 2023.

J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339.

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," in *The Twelfth International Conference on Learning Representations*, 2023.

Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 4083– 4091.

M. Guo, J. Ainslie, D. C. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and

Y. Yang, "Longt5: Efficient text-to-text transformer for long sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 724–736.

A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Ho¨hne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4459–4469.

Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang,

H. Wang *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, p. 32, 2023.

O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 39–48.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le,

Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux,

T. Lacroix, B. Rozie`re, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

W. Khallouli, M. S. Uddin, A. Sousa-Poza, J. Li, and S. Kovacic, "Lever- aging transformer-based ocr model with generative data augmentation for engineering document recognition," *Electronics*, vol. 14, no. 1, p. 5, 2024.

M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 11, 2023, pp. 13 094–13 102.

G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun,

Han, and S. Park, "Ocr-free document understanding transformer (2022)," *URL https://arxiv. org/abs/2111.15664*, 2024.

G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, 2021, pp. 874–880.

J. Zhang, Q. Zhang, B. Wang, L. Ouyang, Z. Wen, Y. Li, K.-H. Chow,

He, and W. Zhang, "Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 17 443–17 453.

D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei,

Nourbakhsh, and X. Liu, "Docllm: A layout-aware generative lan- guage model for multimodal document understanding," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 8529–8548.

C. Luo, Y. Shen, Z. Zhu, Q. Zheng, Z. Yu, and C. Yao, "Layoutllm: Layout instruction tuning with large language models for document un- derstanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 630–15 640.

W. Wang, H. Hu, Z. Zhang, Z. Li, H. Shao, and D. Dahlmeier, "Document intelligence in the era of large language models: A survey," *arXiv preprint arXiv:2510.13366*, 2025.

S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 993–1003.