Beyond Traditional Business Intelligence: Large Language Models as Game-Changers in Document Processing

Tea Krčmar, Dina Šabanović, Miljenko Švarcmajer, Mirko Köhler, Ivica Lukić
J. J. Strossmayer University of Osijek, FERIT Osijek, Croatia
Cara Hadrijana 10b, Osijek
{tea.krcmar, dina.sabanovic, miljenko.svarcmajer, mirko.kohler, ivica.lukic}@ferit.hr

Marko Carević
Uvid d.o.o., Zagreb, Croatia
Hrelička 96, Zagreb
marko@uvid.hr

Abstract. *Unstructured data challenges traditional business intelligence (BI), particularly when processing layout-rich documents like scanned images or PDFs. This paper explores using large language models (LLMs) for extracting structured information, focusing on the LMDX (Language Model-based Document Information Extraction and Localization) approach. LMDX encodes document layout by inserting coordinate tokens directly into the prompt, eliminating the need for vision-based models or architectural changes. We also propose a blockchain-enabled, cloud-based pipeline that integrates LMDX for secure and scalable document intelligence. The paper presents key findings and outlines future research to enhance document processing and automated decision-making in modern BI systems.*

Keywords. Business Intelligence, Document Processing, Information Extraction, Layout Analysis, Machine Learning

1 Introduction

Today, most of the data generated in the digital world is classified as unstructured, characterized by the absence of a predefined model or organized structure, which prevents it from being stored in traditional relational databases. Unstructured data, often referred to as qualitative data, is recognized for its rich information, but is also criticized for its complexity in analysis due to its ambiguous and unstandardized nature (Oza et al., 2023). The exponential growth of unstructured data, mostly scanned images, handwritten notes, PDFs, emails, and text documents, requires intelligent document processing solutions (Mahadevkar et al., 2024; Singh & Hooda, 2023). Although structured data is well managed, difficulties are encountered in processing, retrieving, and understanding information from a variety of document formats by traditional business intelligence (BI) systems (Majid et al., 2024). Through the application of machine learning and natural language processing techniques, unstructured data can be analyzed to extract valuable information (Mahadevkar et al., 2024), which is then used to inform business decisions, enhance customer understanding, and support strategic planning for future needs (Sever, 2024).

Recent advances in deep learning, especially large-scale language models such as one proposed by (Perot et al., 2024) and pre-trained neural networks (Sage et al., 2019, 2021; X. Yang et al., 2017), have redefined how documents are analyzed and structured (Mahadevkar et al., 2024). Tasks such as text recognition, entity extraction, summarization, and multi-modal data processing are excelled in by these models (He et al., 2024; Yue et al., 2024), making them essential to automate processes that were previously required to be handled by considerable human effort. Unlike rule-based or template-based approaches, machine learning (ML) document processing can adapt to different document types, context can be inferred, and accuracy can be increased over time (Baviskar et al., 2021).

The potential of machine learning to transform document processing in business intelligence is explored in this paper. First, a comprehensive analysis of the impact of ML approaches and their use in BI is provided. Secondly, the problem and impact of unstructured data in business intelligence are examined. In the following chapter, LLMs for information extraction from layout-rich documents are explored with emphasis on LMDX by (Perot et al., 2024). Next, we propose a way to integrate LMDX into a cloud-based business intelligence system. Finally, key findings are summarized and future research directions are proposed to address the ongoing challenges of analyzing unstructured documents and their role in business intelligence.

2 Advancing Business Intelligence with Machine Learning

Machine learning has been recognized as a game-changer in business intelligence (Bharadiya, 2023; Sage et al., 2019), with its ability to enable organizations to move beyond traditional, static reporting to dynamic, data-driven decision-making (Hindle et al., 2019). Unlike rule-based systems, which are designed to rely on predetermined logic, ML models are trained on patterns in data, allowing hidden insights, anomalies, and accurate predictions to be uncovered (Bloch & Sacks, 2018; Taddy, 2018). Tasks involving high-dimensional data, where accuracy and efficiency are difficult to maintain with traditional methods, are particularly well-suited to these models (Taddy, 2018). Decision-making processes across all sectors are improved, operations are optimized, customer experience is enhanced, strategic growth is driven, and businesses are given a competitive advantage (Attaran & Deb, 2018; Prasanth et al., 2023). Additionally, new market opportunities and potential supply chain risks can be proactively identified before companies fall victim to predictive analytics, which is regarded as an essential part of machine learning in business intelligence. Rather than responding to problems as they arise, a forward-looking strategy is enabled, allowing proactive action to be taken (Pasupuleti et al., 2024).

An example of ML's capabilities is seen in demand forecasting, a crucial yet intricate aspect of supply chain management, where precise predictions of future demand must be made while accounting for numerous dynamic variables (Abolghasemi et al., 2019; Pasupuleti et al., 2024). Predefined statistical approaches are traditionally used in forecasting models, yet they struggle to capture nuanced market fluctuations, leading to inefficiencies such as overstocking or stockouts (J. Wang et al., 2018). In contrast, machine learning techniques, including ensemble methods and deep learning, are applied to vast and diverse datasets, ranging from historical sales records to external factors like economic indicators, seasonal trends, and consumer sentiment analysis (Pasupuleti et al., 2024). Predictive models are continuously refined with real-time data, significantly improving forecast accuracy and allowing inventory levels to be optimized, waste to be reduced, and resource allocation to be enhanced (Morariu et al., 2020).

Beyond forecasting, logistical operations such as vehicle routing, transportation planning, and inventory optimization are significantly refined through ML applications (Alam et al., 2020; G.-B. Huang et al., 2006). Real-time data sources, including traffic conditions, weather patterns, and customer order updates, are integrated into ML algorithms, allowing delivery schedules to be dynamically adjusted and routes to be optimized, reducing transportation costs and increasing efficiency (Jayaprakash et al., 2021). Through this adaptability, customer demands can be met more effectively, ensuring improved service reliability and satisfaction (Pasupuleti et al., 2024). Additionally, strategic decision-making is enhanced as market trends, shifts in consumer behavior, and supply chain vulnerabilities are uncovered. With these predictive insights, operational adjustments can be made proactively, allowing risks to be mitigated and emerging opportunities to be capitalized upon (M. Yang et al., 2019). Across industries, ML's transformative impact has been demonstrated through case studies, in which AI-driven analytics have been leveraged to streamline supply chain processes, minimize disruptions, and maintain a competitive edge (Pasupuleti et al., 2024).

3 The Issue of Unstructured Data in Business Intelligence

The main problem with unstructured data in BI lies in its inherent complexity (Attaran & Deb, 2018; Majid et al., 2024). Numerous information sources are unstructured or semi-structured, such as customer emails, web pages with competitor information, sales force reports, research paper repositories, and more (Baars & Kemper, 2008). These diverse types of data come in different formats—ranging from free-form text to images and audio(Baviskar et al., 2021)—which complicates the application of traditional BI tools, typically designed to handle structured, tabular data (Baars & Kemper, 2008). Additionally, extracting embedded information from visual data such as scanned documents, pictures and videos poses a challenge to current processing techniques (Adnan & Akbar, 2019). This lack of standardisation is exacerbated by the fact that BI systems rely on predefined models and structured data formats to produce meaningful insights (Sharma et al., 2021).

As the amount of unstructured data increases exponentially, traditional BI systems face considerable constraints in processing data and making strategic decisions. Traditional BI systems, which aggregate and process data in life cycles (Czvetko et al., 2022), fail to provide timely analysis for stratigic planning, and with it limit businesses' from fully capitalising on emerging trends in real-time. The sectors producing large volumes of non-structured content, such as legal contracts or medical records, are prime examples of data-driven decision-making problems (Kumar & Singh, 2019; S.-H. Park et al., 2021). This challenge also extends to businesses, which may miss out on valuable insights contained in customer feedback, market trends, social media content, feedback on news or operational reports. Lack of strategic planning is a significant factor contributing to the failure of businesses, especially those just starting (Chukwuelue, 2024).

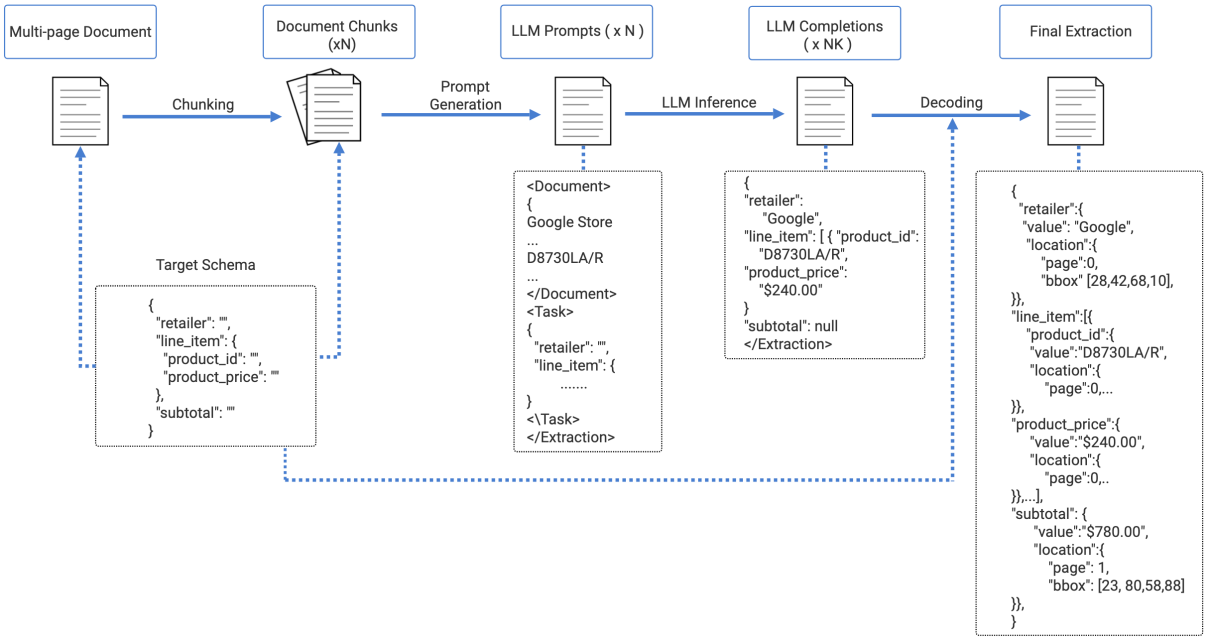

Figure 1. Methodology of LMDX

4 LLMs as Tools for Layout-Rich Document Data Extraction

Within organisations, the key information is often contained in layout-aware documents (LRDs), such as reports and presentations, which combine layout-aware visual elements (e.g., graphs, tables, charts) with text structure and content (Colakoglu et al., 2025; S. Park et al., 2019; Z. Wang et al., 2023; Zmigrod et al., 2024). LRDs are challenging traditional natural language processing (NLP) techniques designed for plain text (Cui et al., 2021; Tang et al., 2023) by incorporating visual and structural features to improve the extraction of information from LRDs (Colakoglu et al., 2025). However, these models require significant fine-tuning of the dataset to be accurate; users must manually annotate the training dataset with bounding boxes and extraction elements for each new document set (Colakoglu et al., 2025).

The emergence of large-scale language models, such as GPT-4 (Achiam et al., 2023), has significantly advanced the NLP domain due to their exceptional ability to understand and generate text (Liu et al., 2024; D. Xu et al., 2023). Retraining language teachers via automatic regression prediction allows them to capture natural patterns and semantic knowledge within the text corpus (Lyu et al., 2024; D. Xu et al., 2023). This enables LLMs to perform zero and few-shot learning, model various tasks consistently, and serve as tools for data augmentation. Additionally, LLMs can act as intelligent agents for complex planning and task execution, leveraging memory mining and various tools to increase efficiency and ensure successful task completion (D. Xu et al., 2023).

Since visually rich documents (VRDs) combine both textual and visual elements, with spatial positioning crucial for understanding, many approaches have explored custom architectures and pretraining strategies to model the relationship between text, layout, and image modalities. For instance, (Y. Xu et al., 2022) use a separate image encoder to add features to token encodings, while (Y. Huang et al., 2022) jointly model page image patches and tokens, using self-supervised pretraining tasks to learn modality connections (Perot et al., 2024). (Hong et al., 2022) propose encoding the relative 2D distances of text blocks in transformer attention and learning from unlabeled documents with an area-masking strategy. (Kim et al., 2022) and (Lee et al., 2023) abandon the text modality entirely, opting for a vision transformer encoder with an autoregressive decoder pretrained on pseudo-optical character recognition (OCR) and region masking tasks. LLMs for extraction have mostly been studied in the text domain (Keraghel et al., 2024), either generally (Laskar et al., 2023) or domain-specific (De Toni et al., 2022; Hu et al., 2023). (S. Wang et al., 2023) use an LLM to insert special tokens to mark the boundaries of target entities, building a layout-aware LLM with various document understanding capabilities (Perot et al., 2024).

Among various approaches and previous research, the LMDX (Language Model-based Document Information Extraction and Localization) framework by Perot et al. (2024) stands out. It reframes information extraction from visually rich documents as a sequence-to-JSON generation task, enabling large language models to perform layout-aware extraction using only text-based prompts. LMDX encodes document layout by appending normalized coordinate tokens to each line or word segment, allowing the LLM to infer

spatial relationships without needing vision encoders or architecture modifications. As shown in Fig.1, the pipeline begins by chunking documents: pages are processed individually and truncated line by line until each chunk satisfies the model's token limit. This approach minimizes entity fragmentation across pages. Each chunk is embedded in a structured prompt that includes the text, coordinate tokens, a task description, and a predefined extraction schema formatted in JSON. This schema supports singular, repeated, and hierarchical entities. During inference, multiple completions are sampled from the LLM for each chunk to introduce diversity. These completions are structured JSON outputs in which each extracted value is paired with its corresponding coordinate tokens. The decoding stage then verifies that the extracted text exists at the specified coordinates in the original document. If the referenced segment does not match exactly—either by content or location—the prediction is flagged as a hallucination and discarded. For valid outputs, bounding boxes are reconstructed by aggregating the positions of all verified segments. Finally, majority voting across completions is used to consolidate predictions, improving both consistency and resilience to spurious outputs. Zero-shot extraction allows processing of new document types without training, while fine-tuning improves accuracy with minimal data. Multiple responses are aggregated for reliability, making LMDX effective for automating document processing tasks. LMDX shows state-of-the-art results with space for improvement.

5 Integrating Data Extraction into Cloud

Integrating LMDX (Language Model-based Document Information Extraction and Localization) into a blockchain-enabled cloud-based business intelligence (BI) system, shown in Fig. 2, presents a novel approach to secure, decentralized, and intelligent document processing. LMDX leverages large language models (LLMs) to extract structured information from visually rich documents while ensuring entity localization. When combined with a blockchain infrastructure, this integration enhances data integrity, auditability, and automation in enterprise workflows.

The proposed architecture incorporates a cloud-hosted LMDX deployment that interacts with a blockchain-based storage and verification system. An automated system processes documents, with LMDX converting unstructured information into structured formats. To ensure tamper-resistant integrity and verifiable audit trails, the processed data is hashed and recorded on a blockchain ledger. The structured outputs are subsequently integrated into a BI framework for advanced analytics, visualization, and decision support. A decentralized storage layer maintains the immutability of raw document data while allowing for controlled access through smart contracts deployed on blockchain platforms such as Ethereum or Hyperledger Fabric.

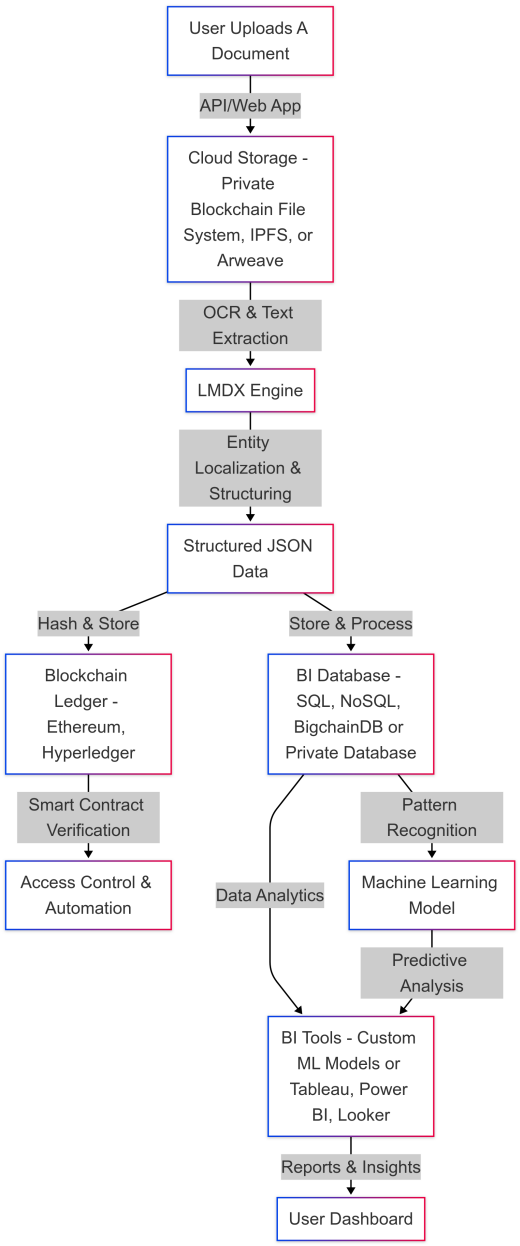

Figure 2. Implementation of LMDX into a business system in the cloud

This integration enables a secure, scalable, and automated workflow f or d ocument i ntelligence, particularly in applications such as financial a uditing, regulatory compliance, and enterprise contract analysis. By utilizing machine learning for data extraction and blockchain for authentication, companies can boost operational productivity, minimize fraud potential, and maintain compliance with legal standards such as GDPR and financial reporting s tandards. The B I system further enriches decision-making processes by providing real-time insights, trend analysis, and anomaly

detection based on structured document data. The synergy between LLM-based document intelligence and blockchain's decentralized trust framework establishes a robust, future-proof solution for enterprises seeking to streamline document-driven workflows in a secure manner.

6 Conclusion and Future Work

Machine learning, in particular through large-scale language models, has proved to be indispensable to overcome the limitations of traditional BI systems in the processing of unstructured data. By automating the extraction of information from structured documents, ML improves the accuracy of data analysis, reduces human effort and allows for real-time decision-making based on data. Although the best available techniques are showing promising results, further improvements in fine-tuning and model adaptability are needed to further optimise their use in a variety of document types. Future research will focus on using LMDX (Language Model-based Document Information Extraction and Localization) as a foundation for further advancing document extraction techniques. We plan to enhance its capabilities, refining the model for better accuracy and adaptability to various document types. Additionally, we aim to integrate this improved version of LMDX into cloud-based business intelligence systems, enabling real-time, scalable document processing. This integration will help streamline BI processes and empower organizations with more efficient, automated decision-making capabilities in the cloud.

Acknowledgments

This work was conducted as part of the project *Research of Advanced Algorithms and Solutions for Innovative Business Intelligence in the Cloud* - NPOO.C3.2.R3-I1.04.0128.

References

Abolghasemi, M., Gerlach, R., Tarr, G., & Beh, E. (2019). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. https://doi.org/10.48550/arXiv.1909.13084

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H.-i., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2023). Gpt-4 technical report. https://api.semanticscholar.org/CorpusID:257532815

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*. https://doi.org/10.1177/1847979019890771

Alam, M., Alam, M., Roman, M., Tufail, M., Khan, U., & Khan, M. (2020). Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture, 273–280. https://doi.org/10.1109/ICEEE49618.2020.9102505

Attaran, M., & Deb, P. (2018). Machine learning: The new 'big thing' for competitive advantage. *5*, 277–305. https://doi.org/10.1504/IJKEDM.2018.10015621

Baars, H., & Kemper, H.-G. (2008). Management support with structured and unstructured data - an integrated business intelligence framework. information systems management 25(2):132-148.doi: 10.1080/10580530801941058. *IS Management*, *25*, 132–148. https://doi.org/10.1080/10580530801941058

Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2021.3072900

Bharadiya, J. P. (2023). The role of machine learning in transforming business intelligence. *Int. J. Comput. Artif. Intell.*, *4*(1), 16–24. https://doi.org/10.33545/27076571.2023.v4.i1a.60

Bloch, T., & Sacks, R. (2018). Comparing machine learning and rule-based inferencing for semantic enrichment of bim models. *Automation in Construction*, *91*, 256–272. https://doi.org/10.1016/j.autcon.2018.03.018

Chukwuelue, A. (2024). Surpassing $1to3$ million revenue threshold: Analyzing why small businesses miss the mark. *Journal of Technology and Systems*, *6*, 14–38. https://doi.org/10.47941/jts.2037

Colakoglu, G., Solmaz, G., & Fürst, J. (2025). Problem solved? information extraction design space for layout-rich documents using llms. https://arxiv.org/abs/2502.18179

Czvetko, T., Kummer, A., Ruppert, T., & Abonyi, J. (2022). Data-driven business process management-based development of industry 4.0 solutions. *CIRP Journal of Manufacturing Science and Technology*, *36*, 117–132. https://doi.org/10.1016/j.cirpj.2021.12.002

De Toni, F., Akiki, C., De La Rosa, J., Fourrier, C., Manjavacas, E., Schweter, S., & Van Strien, D. (2022, May). Entities, dates, and languages: Zero-shot on historical texts with t0. In A. Fan, S. Ilic, T. Wolf, & M. Gallé (Eds.), *Proceedings of bigscience episode #5 – workshop on challenges & perspectives in creating large language models*

(pp. 75–83). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.7

He, X., Zhou, M., Xu, X., Ma, X., Ding, R., Du, L., Gao, Y., Jia, R., Xu, C., Han, S., Yuan, Z., & Zhang, D. (2024). Text2analysis: A benchmark of table question answering with advanced data analysis and unclear queries. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 18206–18215. https://doi.org/10.1609/aaai.v38i16.29779

Hindle, G., Kunc, M., Mortenson, M., Oztekin, A., & Vidgen, R. (2019). Business analytics: Defining the field and identifying a research agenda. *European Journal of Operational Research*, *281*. https://doi.org/10.1016/j.ejor.2019.10.001

Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., & Park, S. (2022). Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. https://arxiv.org/abs/2108.04539

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., Li, Y., Li, J., Jiang, X., & Xu, H. (2023). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association : JAMIA*, *31*, 1812–1820. https://api.semanticscholar.org/CorpusID:257805032

Huang, G.-B., Zhu, Q.-Y., & Siew, C. (2006). Real-time learning capability of neural networks. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *17*, 863–78. https://doi.org/10.1109/TNN.2006.875974

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 4083–4091. https://doi.org/10.1145/3503161.3548112

Jayaprakash, S., Nagarajan, M. D., Prado, R. P. d., Subramanian, S., & Divakarachari, P. B. (2021). A systematic review of energy management strategies for resource allocation in the cloud: Clustering, optimization and machine learning. *Energies*, *14*(17), 5322.

Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. https://api.semanticscholar.org/CorpusID:267060999

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022, October). Ocr-free document understanding transformer. https://doi.org/10.1007/978-3-031-19815-1_29

Kumar, S., & Singh, M. (2019). Big data analytics for healthcare industry: Impact, applications, and tools. *Big Data Mining and Analytics*, *2*(1), 48–57. https://doi.org/10.26599/BDMA.2018.9020031

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S. R., & Huang, J. (2023). A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *ArXiv*, *abs/2305.18486*. https://api.semanticscholar.org/CorpusID:258967462

Lee, K., Joshi, M., Turc, I., Hu, H., Liu, F., Eisenschlos, J., Khandelwal, U., Shaw, P., Chang, M.-W., & Toutanova, K. (2023). Pix2struct: Screenshot parsing as pretraining for visual language understanding. https://arxiv.org/abs/2210.03347

Liu, Q., He, Y., Xu, T., Lian, D., Liu, C., Zheng, Z., & Chen, E. (2024). Unimel: A unified framework for multimodal entity linking with large language models, 1909–1919. https://doi.org/10.1145/3627673.3679793

Lyu, Y., Niu, Z., Xie, Z., Zhang, C., Xu, T., Wang, Y., & Chen, E. (2024). Retrieve-plan-generation: An iterative planning and answering framework for knowledge-intensive llm generation, 4683–4702. https://doi.org/10.18653/v1/2024.emnlp-main.270

Mahadevkar, S., Patil, S., Kotecha, K., Soong, L., & Choudhury, T. (2024). Exploring ai-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, *11*. https://doi.org/10.1186/s40537-024-00948-z

Majid, M., Marinova, D., Hossain, A., & Rummani, F. (2024). Use of conventional business intelligence (bi) systems as the future of big data analysis. *American Journal of Information Systems*, *9*, 1–10. https://doi.org/10.12691/ajis-9-1-1

Morariu, C., Morariu, O., Raileanu, S., & Borangiu, T. (2020). Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems. *Computers in Industry*, *120*, 103244. https://doi.org/10.1016/j.compind.2020.103244

Oza, R., Punjani, D., & Domadiya, D. (2023). Analysis of unstructured data using artificial intelligence. *International Journal of Creative Research Thoughts*, *11*(5), 2320–2882.

Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2019). Cord: A consolidated receipt dataset for post-ocr parsing. https://api.semanticscholar.org/CorpusID:207900784

Park, S.-H., Lee, D.-G., Park, J.-S., & Kim, J. W. (2021). A survey of research on data analytics-based legal tech. *Sustainability*, *13*, 8085. https://doi.org/10.3390/su13148085

Pasupuleti, V., Thuraka, B., Kodete, C. S., & Malisetty, S. (2024). Enhancing supply chain agility and sustainability through machine learning: Optimization techniques for logistics and inventory management. *Logistics*, *8*, 73. https://doi.org/10.3390/logistics8030073

Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R. S., Wang, Z., Wang, Z., Mu, J., Zhang, H., Lee, C.-Y., & Hua, N. (2024). LMDX: Language model-based document information extraction and

localization. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 15140–15168). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.899

Prasanth, A., Vadakkan, D. J., Surendran, P., & Thomas, B. (2023). Role of artificial intelligence and business decision making. *International Journal of Advanced Computer Science and Applications*, *14*(6). https://doi.org/10.14569/IJACSA.2023.01406103

Sage, C., Aussem, A., Elghazel, H., Eglin, V., & Espinas, J. (2019). Recurrent neural network approach for table field extraction in business documents. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1308–1313. https://doi.org/10.1109/ICDAR.2019.00211

Sage, C., Douzon, T., Aussem, A., Eglin, V., Elghazel, H., Duffner, S., Garcia, C., & Espinas, J. (2021). Data-efficient information extraction from documents with pre-trained language models. *ICDAR Workshops*. https://api.semanticscholar.org/CorpusID:237345634

Sever, M. M. (2024). Digital transformation and business intelligence (bi) in the industry 4.0 (i 4.0) age. In P. Keikhosrokiani (Ed.), *Data-driven business intelligence systems for socio-technical organizations* (pp. 28–54). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1210-0.ch002

Sharma, K., Shetty, A., Jain, A., & Dhanare, R. (2021). A comparative analysis on various business intelligence (bi), data science and data analytics tools, 1–11. https://doi.org/10.1109/ICCCI50826.2021.9402226

Singh, S., & Hooda, S. (2023). A study of challenges and limitations to applying machine learning to highly unstructured data. *7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–6. https://doi.org/10.1109/ICCUBEA58933.2023.10392115

Taddy, M. (2018, February). *The Technological Elements of Artificial Intelligence* (NBER Working Papers No. 24301). National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/24301.html

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*. https://api.semanticscholar.org/CorpusID:117248575

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). Gpt-ner: Named entity recognition via large language models. *ArXiv*, *abs/2304.10428*. https://api.semanticscholar.org/CorpusID:258236561

Wang, Z., Zhou, Y., Wei, W., Lee, C.-Y., & Tata, S. (2023). Vrdu: A benchmark for visually-rich document understanding. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5184–5193. https://doi.org/10.1145/3580305.3599929

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., & Chen, E. (2023). Large language models for generative information extraction: A survey. *Frontiers Comput. Sci.*, *18*, 186357. https://api.semanticscholar.org/CorpusID:266690657

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., & Zhou, L. (2022). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. https://arxiv.org/abs/2012.14740

Yang, M., Lim, M. K., Qu, Y., Ni, D., & Xiao, Z. (2019). Supply chain risk management with machine learning technology: A literature review and future research directions. *Computers & Industrial Engineering*, *175*, 108859.

Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., & Lee Giles, C. (2017). Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue, C., Xu, X., Ma, X., Du, L., Ding, Z., Han, S., Zhang, D., & Zhang, Q. (2024). Extract information from hybrid long documents leveraging llms: A framework and dataset. https://doi.org/10.48550/arXiv.2412.20072

Zmigrod, R., Wang, D., Sibue, M., Pei, Y., Babkin, P., Brugere, I., Liu, X., Navarro, N., Papadimitriou, A., Watson, W., Ma, Z., Nourbakhsh, A., & Shah, S. (2024). Buddie: A business document dataset for multi-task information extraction. https://arxiv.org/abs/2404.04003