

A Supervised Framework for Document Processing at Scale with Large Language Models in Credit-Risk Research

Cosmin COJOCARU*

National University of Science and Technology Politehnica Bucharest, Bucharest, Romania

ORCID ID: 0009-0004-2255-110X

*Corresponding author, e-mail address: cosmin.cojocaru1906@stud.faima.upb.ro

Sorin IONESCU

National University of Science and Technology Politehnica Bucharest, Bucharest, Romania

e-mail address: sc.ionescu@gmail.com

Abstract. In a time of exponentially growing unstructured digital content, automatic information extraction from huge corpora is an indispensable ability for the modern economy and research. Credit risk research is based on a wide scientific literature content, hard to be synthesized and compared because of its multiple layers of models and results. Recent breakthroughs in large language models mitigate this limitation for partially automated extraction, but the majority of tools available today suffer from low reproducibility, accuracy, and faithfulness of extracted information. This paper presents an approach of a supervised large language model to generate semi-structured data used for performing automatic document processing on credit risk research papers. This work challenges how well a supervised approach can extract structured information, such as model names, performance measures, validation types and reference lists from diverse sources and digital formats. The method uses schema based guidance with automatic text retrieval. Documents are processed independently and transformed into a structured schema that is compared with a ground truth to evaluate accuracy, completeness, and faithfulness. The key evaluation parameters are title, authors, year, DOI, journal, model identification, metrics and references. Results demonstrate that the proposed supervised strategy is capable of producing higher precision, completeness and reproducibility than using general large language model extractions. It demonstrates much lower hallucination and omission error and increased content coverage. The framework is able to parse a variety of academic papers, and provides robust results across required metadata. Outside of the scientific literature, the architecture is applicable to other types of unstructured documents to credit risk research, such as annual financial statements, audit reports or legal verdicts. These results validate that a supervised and tractable pipeline forges the basis of large scale evidence mapping and analytic automation, which enable knowledge integration and future predictive modeling in risk analytics and beyond.

Keywords: supervised document processing; large language models; credit-risk research; metadata extraction; academic text analysis; model validation; knowledge management

How to cite: COJOCARU, C., & IONESCU, S. . A Supervised Framework for Document Processing at Scale with Large Language Models in Credit-Risk Research. *International Conference of Management and Industrial Engineering*, 12. <https://doi.org/10.56177/12icmie2025.196>

Introduction

Increasing adoption of digital research has led to a vast and accelerating amount of scientific papers, available in PDF format. The main disadvantage of this layout is that it does not support easy analysis and extraction of data at scale. For scientific research communities interested in aggregating and integrating the output of many studies, automatic extraction of structured information such as models, validation methods, datasets, and references is increasingly necessary to ensure transparency and reproducibility.

The latest large language models (LLM) have achieved the state of art in text understanding; however, direct application to structured data extraction, especially for large documents is still below expectations. In an unsupervised approach, the results are incomplete or incoherent. Models may even hallucinate information not present in the source text. These mistakes decrease the precision and ability to automate metadata extraction. Those

constraints are crucial in certain fields such as credit risk research, where models need to be transparent and traceable, considering they are used for credit risk decisions.

In this study, we outline a supervised framework to extract specific metadata from scientific papers. It integrates document parsing, optical character recognition (OCR), table recognition and text extraction with the help of large language models. Every paper is parsed and checked for completeness and internal consistence. The method applied locally for data extraction, and produces traceable outputs that can be further analyzed. The central question to be addressed is whether a supervised approach like this can produce more accurate and consistent metadata than other available models. For this, two hypotheses are tested:

H1: *Schema constraints and validation through supervision help in accuracy and completeness.*

H2: *The supervised approach is more reliable and readiness for automation.*

The research uses and tests these questions on a database of 20 credit risk papers. The aim is to demonstrate that reliable automated document processing does not depend on model scale, but rather on structured supervision. The presented framework also provides a basis for scalable and auditable extraction of information from other non-structured sources that can be used in credit risk modeling, such as annual financial statements or court decisions.

Literature review

Document understanding and metadata extraction

Structured information extraction has been a long-standing problem in information science. Early systems like GROBID [1] and CERMINE [2] employed rules and basic machine learning to recognize fields like title, author or references. Both systems made use of segmentation rules in conjunction with classifiers trained on annotated training samples. This was a preliminary attempt in reproducible metadata extraction. But in presence of more complex page layout and various columns, tables or figures, this was not sufficient.

Subsequent studies proposed models that take both the text and layout information. The position where each word appears on the page was also employed, inspired by LayoutLM [3] as they considered how words are determined in a textual context and on a document level. This contributed to the enhancement of entities and text block recognition. Donut [4] took the step forward by eliminating the requirement to perform a separate optical character recognition stage. It could directly process document, images, extract text and structure at the same time. TrOCR [5] merged the visual and textual steps into one by using a single model for text recognition. Research on table extraction had a similar path. The models presented by the Table Transformer [6] trained on the PubTables-1M dataset, helped identifying table boundaries and cells in scientific papers. This was a more precise recovery of numeric data and better understanding of document structure.

These models jointly constitute the current base for document parsing. However, post processing is still required, in order to correct errors and extract data from different sources into a unified schema.

Reliability challenges in LLM-based extraction

Big language models have incorporated contextual reasoning while processing documents. They are able to infer text meaning and connect ideas between sections. They also tend to generate untrustworthy outputs when they are used without supervision. Sahoo et al. [7] demonstrated that the models tend to change field names or add text that is not present in the original document. Manakul et al. [8] discovered that errors persist even after models validate their output. Such a practice inhibits reproducibility and trust.

Reliability can be understood through three simple dimensions. The first is validity, which means the output follows the expected structure. The second is faithfulness, which measures how well the extracted information matches the source text. The third is completeness, which shows whether all required elements are present. Many chat based systems perform unevenly across these dimensions. They often lose structure when the document is long or when its layout interrupts the logical flow of text. Approaches like SelfCheckGPT [8] improve internal consistency but do not ensure that fields are syntactically correct or complete. These gaps show the need for supervised frameworks that use fixed schemas and validation steps to check results before acceptance.

Credit risk research as test domain

Credit risk research provides a challenging testing ground for metadata extraction algorithms. Previous research in this area recognized statistical methods like the Altman Z-score [9] and the Ohlson logit model [10]. Recent research has investigated a wide variety of machine learning, e.g., neural networks, support vector machines, and ensemble classifiers to determine credit risk [11]. Articles in this area tend to contrast a number of different models, outline

intricate validation procedures, and include extensive tables comprising numerical outcomes. They also include extensive reference lists. This sort of form and diversity is most suitable for the estimation of extraction reliability.

At present, no comprehensive study has compared supervised language model extraction with general chat based systems on a predetermined schema. Current tools extract information but are not able to return an aggregate rating of accuracy, completeness, and consistency in controlled conditions. Missing is an auditable system that integrates deterministic preprocessing with schema extraction defined and quantifiable reliability scores.

The current study addresses this gap. It introduces a supervised and local framework that applies schema rules and validation metrics to scientific papers in credit risk research. The framework transforms each document into a structured representation and evaluates the quality of extracted data. By grounding the process in measurable reliability criteria, this approach supports reproducible and transparent document processing. It also lays the foundation for extending supervised extraction to other large unstructured sources used in credit risk analysis, including financial statements and legal documents.

Methodology

The methodological framework of this study was built to test the main hypotheses. These hypotheses state that supervision through a fixed schema increases the accuracy, completeness, and stability of metadata extraction when compared with general chat based language models. To test this idea, a comparative experiment was carried out. Both the supervised framework and several chat based systems processed the same collection of credit risk research papers under identical schema conditions.

The research design followed four main phases. These were document preprocessing and parsing, metadata extraction on defined schema, validation and caching of results, and finally, evaluation. Each phase is described in the following sections.

Document preprocessing and parsing

The document collection included a variaton of papers on credit risk research. It covered both classical statistical studies such as Altman [8] and Ohlson [9], but also more recent machine learning models such as Barboza et al. [10]. Each paper was available in PDF format and was processed through a Python based text extraction engine. The engine read page geometry, text blocks, and embedded figures. For image based pages, optical character recognition was applied using open source software to recover text.

Tables were detected through a transformer model trained on the PubTables 1M dataset [6]. The model identified table areas and cells on each page. This produced a structured layout that could be used for the next stage of semantic extraction. The result of this step was a consistent digital representation of each paper that preserved text order, figures, and tables in a clean structure.

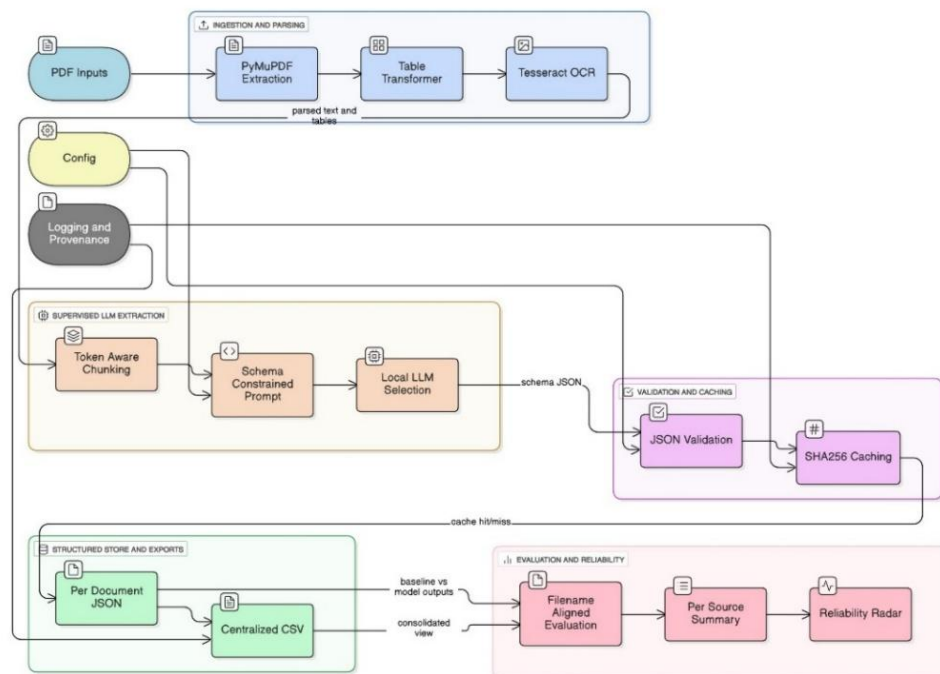


Figure 1: Architecture of the supervised LLM-based metadata-extraction framework.
Source: Authors' own research.

Metadata extraction based on defined schema

In the second phase, the parsed text was converted into structured metadata following a fixed schema. Each document was divided into overlapping text segments to keep meaning continuity within the model context. The supervised prompt specified the schema fields and the required output format. The schema included title, authors, year, journal or conference, doi, models, metrics, validation types, explainability methods, datasets, and references.

Partial results were merged deterministically through a Map Reduce procedure. This ensured that overlapping content was reconciled and that consistency was preserved for each field. The method prevented schema drift and maintained consistency across all parts of a document. The extraction process used the OpenAI API service in a controlled mode but can be also extended and compatible with local models such as LLaMA or Mistral.

Validation and caching of retrieved data

All output JSONs were validated by two validation processes. The structure validation process tested that the structure conformed to JSON standards. The logical validation process verified that all the required fields existed and were logically valid.

For ensuring non-redundancy, the entire source files were hashed with SHA-256. This made it possible to trace completely from the root PDF to all subsequent results. The process made the pipeline deterministic and would always produce exactly the same copies given the same inputs.

Evaluation of results

The evaluation compared the supervised framework with standard chat based systems and agent style models. Each model processed the same set of papers and produced data that was aligned by filename to ensure one to one comparison. Three main groups of metrics were used: *metadata block* (equality of title, year, doi, and journal_or_conference), *models block* (overlap of model sets, models / metric pairs, and validation types) and *references block* (precision, recall, F1-score, and reference-count accuracy computed from author/year pairs).

Each block is scored on a 0–100 scale, and results are aggregated into a Weighted Score using the formula

$$\text{Weighted score} = 0.35 M_{\{\text{meta}\}} + 0.30 M_{\{\text{models}\}} + 0.35 M_{\{\text{refs}\}}$$

where M_{meta} , M_{models} , and M_{refs} represent the normalized block means. This composite index balances factual identity and referential integrity, both essential for automation readiness.

System level averages and standard deviations are reported in Table 1. Detailed analyses at field and reliability level are presented in Tables 2 and 3.

Software implementation

The framework was developed in Python with a Gradio interface. It uses the libraries gradio, openai, pymupdf, pymupdf4llm, transformers, torch, pandas, numpy, Pillow, tiktoken, openpyxl, and pytesseract [12]. PyMuPDF and pymupdf4llm handle parsing and conversion to text and markdown. Tables are detected with the Microsoft Table Transformer model, while OCR is applied when images or scanned pages are found. Documents are processed either as a single batch or in overlapping chunks estimated through token counts. Each chunk is extracted using a JSON schema and merged with a Map Reduce step. After processing, each file is hashed with SHA-256 for traceability and caching. For metadata extraction, OpenAI API with ChatGPT-5 is used [13] with outputs saved as JSON and CSV.

The comparative evaluation included several mainstream LLM chat systems. Each system processed the same corpus and was instructed to produce metadata conforming to the identical schema. A manually curated baseline JSON served as reference truth for performance scoring.

This integrated design ensures that all observed performance differences result from the supervision mechanism rather than dataset or model variability. By coupling deterministic preprocessing, schema-constrained generation, and quantitative benchmarking, the methodology establishes a transparent and auditable foundation for the results and further discussion.

The full implementation, including the supervised document processing framework, evaluation notebooks, and derived datasets, is publicly available on GitHub [14].

Results and further discussions

Framework evaluation

The supervised framework served as the baseline for all comparisons. Its role was to verify whether schema supervision can ensure reliable metadata extraction across credit risk research papers.

The approach produced stable and transparent outputs with no hallucinations. All retrieved fields matched the actual content of the documents. Unclear data was left empty instead of generating false values. All models and most performance metrics were extracted correctly. Still, in complex tables, only the final summary metrics were extracted, not the intermediate ones. This behavior can be improved through prompt tuning. Bibliographic information was recovered accurately, except for one paper where only 40 out of 148 references were extracted. For scanned low-quality PDFs, such as the Ohlson paper [10], the system recognized OCR limitations and recorded them in the notes field. It handled landscape pages and content correctly and avoided creating artificial data.

Overall, the framework maintained factual accuracy, avoided hallucinations, and offered reproducible, auditable results, confirming its value as a reliable baseline for evaluating language model extraction systems.

Quantitative evaluation summary

Table 1 presents the system-level comparison of the proposed framework to other large language models. Evaluation shows that Claude Sonnet 4.5 and Grok 4 produced partial or fragmented results often requiring multiple prompts, while Gemini 2.5 and NotebookLM failed to return complete structured outputs.

Table 1: Quantitative Evaluation Summary (per system).

Model / Source	N Papers	Weighted Avg (%)	SD (%)	Min (%)	Max (%)
grok4	19	54.8	18.9	35.8	73.7
chatgpt5_agent	20	52.7	12.9	39.8	65.6
chatgpt5_chatllm	4	49.5	10.1	39.4	59.7
gemini2_5_platform	20	45.5	15.8	29.7	61.3
notebooklm	19	41.1	12.0	29.1	53.1
claude_sonnet_4_5	20	31.8	15.1	16.7	46.9

Source: Authors' own research.

These results, graphically illustrated also in Figure 2 confirm that supervised extraction achieves markedly higher reliability than usual conversational assistants or agents. The reduction in standard deviation reflects improved

stability across documents, supporting H1: supervision enhances factual accuracy and completeness. Prior document-understanding research [3], [6] similarly emphasized that structural supervision mitigates variability.

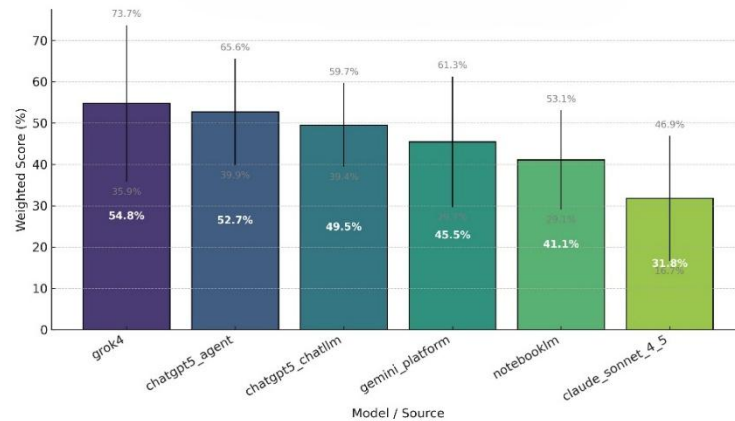


Figure 2: Comparison of LLM systems by weighted score and error dispersion.
Source: Authors' own research.

Field-level performance analysis

Field level metrics presented in Table 2 and as heatmap in Figure 3 show where differences in performance appear between systems. Chat based language models show a poor output across the board. The 100% score for chatgpt-5 on title, DOI and journal are on the back of only 4 papers extracted. Furthermore, considering the complexity and length of the references, here the comparison shows the poorest results, with only Grok-4 managing to reach 70% vs. the baseline (however, in some cases with hallucinated titles included in references).

Table 2: Field-level Accuracy by Semantic Block

Model / Source	Title (%)	Year (%)	DOI (%)	Journal (%)	Models (%)	Metrics (%)	Valid. Ovr. (%)	Ref. F1 (%)	Ref Acc.(%)
grok4	89.5	84.2	57.9	79.0	11.0	22.1	21.0	62.6	69.3
chatgpt5_agent	70.0	90.0	95.0	80.0	42.9	20.2	30.4	32.7	34.0
chatgpt5_chatlm	100.0	75.0	100.0	100.0	25.4	25.0	8.3	18.4	12.3
gemini2_5_platform	90.0	85.0	50.0	75.0	15.7	20.0	21.7	29.3	33.9
notebooklm	73.7	84.2	73.7	73.7	28.2	22.1	7.0	11.9	7.9
claude_sonnet_4_5	75.0	45.0	40.0	50.0	26.6	20.4	15.8	11.5	31.2

Source: Authors' own research.

These results align with previous findings [7], [8], which observed degradation in chat-based LLM outputs when task supervision decreases. The supervised approach confirms that prompt determinism and structured validation outperform unsupervised reasoning for complex academic documents with possible conclusions on any type of other similar document with multiple tables or consistent content blocks.

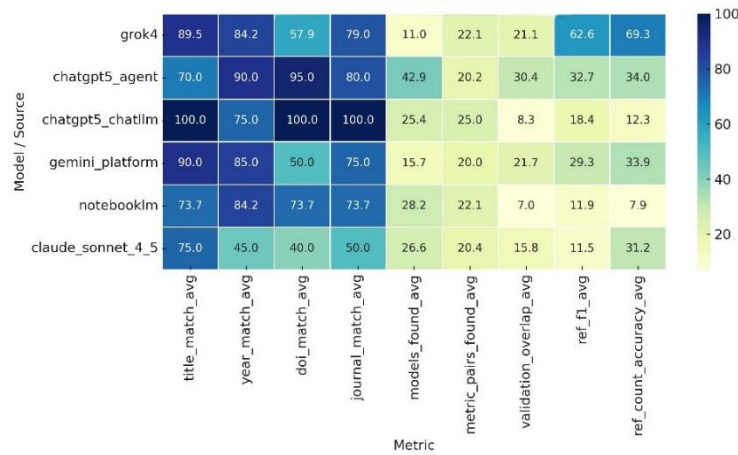


Figure 3: Field-Level Accuracy Heatmap. Darker tones indicate higher reliability.
Source: Authors' own research.

Reliability dimensions

To translate accuracy into interpretable quality indicators, the study introduces five reliability dimensions: JSON Validity, Completeness, Faithfulness, Reference F1, and Ease of Automation.

Table 3: Quantitative Reliability Dimensions

Model / Source	JSON Validity (%)	Completeness (%)	Faithfulness (%)	Reference F1 (%)	Ease of Automation (%)
grok4	60	64.7	43.9	62.6	62.3
chatgpt5_agent	100	74.0	38.5	32.7	86.0
chatgpt5_chatlm	0	87.5	50.0	18.4	0.0
gemini2_5_platform	100	56.5	45.5	29.3	75.2
notebooklm	100	73.7	41.7	11.9	85.8
claude_sonnet_4_5	100	86.0	22.8	11.5	92.7

Source: Authors' own research.

The reliability radar in Figure 4 shows the balance of performance across the five reliability dimensions. The unsupervised approach shows poor faithfulness and F1, with irregular geometries on Validity and Completeness.

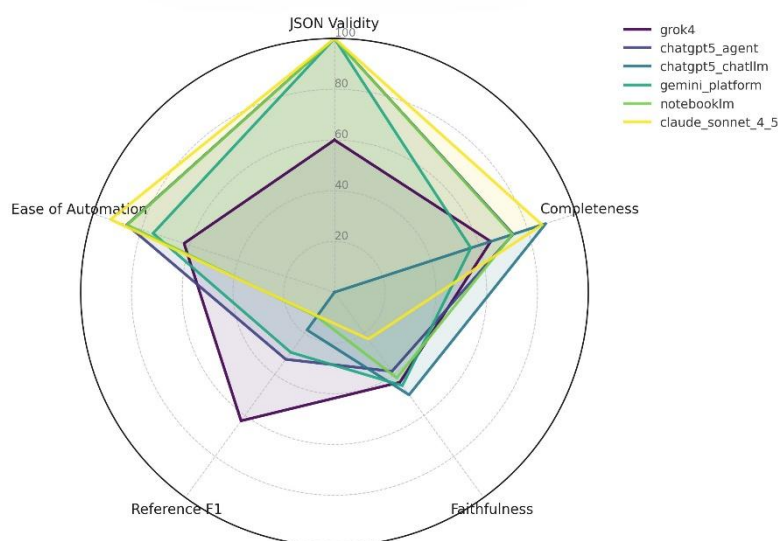


Figure 4. Reliability radar of evaluated systems.
Source: Authors' own research.

Conclusion

This research formulated methods to develop and evaluate a supervised model using large language models for reliable document processing in credit risk studies. The model included controlled supervision, reliable deterministic preprocessing, and schema checking to extract trustworthy metadata from scientific documents. It served as a baseline with these entities compared against various widely used language models showing superiority over the three aspects of precision, completeness, and consistency.

The findings validate that supervision instead of model capacity dictates reliability in automatic data extraction. Claude Sonnet 4.5 and Gemini 2.5 models are good at reasoning ability but poor at structural control for reproducible output. The supervised model secures that each extracted field comes under a tight schema, minimizing hallucinations and having excellent traceability. It performed remarkably well on all axes it was tested on, e.g., JSON validity, completeness, faithfulness, and reference accuracy.

In practice, the framework implements an open and reproducible way of converting unstructured documents into structured datasets that can be used for various purposes. It is locally executable, independent of third-party cloud-based services, thus making it viable for regulatory or sensitive environments. The resultant standardized outputs can be used for comparison of new extraction models or for structuring knowledge across organizations.

Although the model is good at handling contemporary scientific articles, it does have some limitations. This version was limited to English language publications, doesn't yet extract all numeric data out of complicated tables (tractable by schema definition), and OCR quality continues to have impact in older or scanned documents. Reference extraction is also probably impacted when citations are spread across pages instead of being gathered at the end. These problems can be alleviated by improved prompting, table splitting, and active learning feedback.

In the future, the same guided system may also be applied beyond academic research papers. Most credit risk reports, including annual financial reports, audit reports, or court verdicts, share the same issues: large documents with difficult to extract meaningful, structured content. If the same schema-based extraction and validation rules are employed, they can also be converted to structured data sets and uncover a totally new field of building predictive models from unstructured data.

Summarily, this research presents both a methodology and a standard for trustworthy document understanding. It demonstrates the manner in which big language models, guided by explicit structure and validation, can attain scalable accurate and reproducible knowledge extraction. The strategy provides a viable basis for clear, auditable, and domain movable AI frameworks in credit risk evaluation and beyond.

The supervised metadata-extraction framework, evaluation scripts, and reproducibility materials are available at *GitHub*: <https://github.com/cojocarucosmin/supervised-llm-document-framework> (release v1.0). The

repository includes all JSON and evaluation files required for reproducibility. Original PDFs are not redistributed due to copyright restrictions.

References

- [1] Lopez, P.: GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, pp. 473–474, ISBN 978-3-642-04345-1, Corfu, Greece, September 2009, Springer, Berlin, (2009).
- [2] Tkaczyk, D. et al.: CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature, *International Journal on Document Analysis and Recognition (IJDAR)*, 18 (2015) 4, pp. 317–335, ISSN 1433-2825.
- [3] Xu, Y. et al.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, pp. 1192–1200, ISBN 978-1-4503-7998-4, virtual event, August 2020, ACM, New York, (2020).
- [4] Kim, G. et al.: OCR-Free Document Understanding Transformer (Donut), *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, pp. 498–517, ISBN 978-3-031-19814-4, Tel Aviv, Israel, October 2022, Springer, Berlin, (2022).
- [5] Li, M. et al.: TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, Available from <https://arxiv.org/abs/2109.10282> Accessed: 2025-09-30.
- [6] Smock, B. et al.: PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 4634–4642, ISBN 978-1-6654-6946-3, New Orleans, USA, June 2022, IEEE, (2022).
- [7] Sahoo, P. et al.: A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11709–11724, ISBN 979-8-89176-168-1, Miami, USA, November 2024, Association for Computational Linguistics, (2024).
- [8] Manakul, P. et al.: SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, Available from <https://arxiv.org/abs/2303.08896> Accessed: 2025-09-30.
- [9] Altman, E. I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, 23 (1968) 4, pp. 589–609, ISSN 0022-1082.
- [10] Ohlson, J. A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 18 (1980) 1, pp. 109–131, ISSN 0021-8456.
- [11] Barboza, F. et al.: Machine Learning Models and Bankruptcy Prediction, *Expert Systems with Applications*, 83 (2017), pp. 405–417, ISSN 0957-4174.
- [12] Smith, R.: An Overview of the Tesseract OCR Engine, *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, pp. 629–633, ISBN 978-0-7695-2822-9, Curitiba, Brazil, September 2007, IEEE Computer Society, (2007).
- [13] OpenAI: ChatGPT-5 Technical Overview, Available from <https://openai.com/research> Accessed: 2025-09-30.
- [14] Cojocaru, C.; Ionescu, S.: Code repository and Colab notebook for “A Supervised Framework for Document Processing at Scale with Large Language Models in Credit-Risk Research”, Available from <https://github.com/cojocarucosmin/supervised-llm-document-framework> Accessed: 2025-10-09.