Berliner Hochschule für Technik

Extraction of tabular data from annual reports with LLMs

Using in context learning with open source models and RAG

submitted by

Simon Schäfer

Matr.-Nr.: 944 521

Department VI – Informatics and Media
Berliner Hochschule für Technik Berlin
presented Master Thesis
to aquiere the academic degree

Master of Science (M.Sc.)

in the field of

Data Science

Date of submission September 1, 2025

Studiere Zukunft

Gutachter

Prof. Dr. Alexander Löser	Berliner Hochschule für Technik
Prof. Dr. Felix Gers	Berliner Hochschule für Technik

Abstract

Content of this thesis is a benchmark on information extraction from PDFs. The focus are annual reports of German companies. Special characteristic of the task is handling hierarchies in tables with financial data to prepare the data for import into a relational database.

The benchmark is composed of three sub tasks and the performance of different open source large language models is tested with different prompting approaches and compared to alternative methods.

Zusammenfassung

Gegenstand dieser Arbeit ist ein Benchmark zur Informationsextraktion aus PDF-Dateien. Dabei wird sich auf das Auslesen der Bilanzen und Gewinn- und Verlustrechnungen aus Jahresabschlüssen deutscher Unternehmen beschränkt. Ein besonderer Aspekt der Aufgabe ist die Berücksichtigung der Hierarchie innerhalb der Tabellen, um die Werte einem festen Schema zuzuordnen und so den Import in eine relationale Datenbank vorzubereiten.

Notes

- Qwen 2.5 hat zweiseitige GuV von IBB entdeckt und zur Anpassung der Ground Truth

Contents

Chapter 1

Introduction

1.1 Motivation

- market: public administration, companies with data of special requirements for treating (secret and personal data (high risk data)) <– DSGVO, AI act
  - next market for hyper scalers might be public administration with local computing clusters
- whom is it helping
- why now: digital sovereignity, AI act; people want NLP AI products, frameworks get easier
- is the problem easier solvable then years ago? why?

1.2 Objectives

- special part of big problem? central question
- two sentences: why this problem? new problem or just a part in the big task? hard to solve of straight forward? research or application? what was not done and why?
- building a system? what task to solve? core functionality? typical use cases?

1.3 Methodology (1 p)

- how to solve the problem?
- what foundations to have in mind?
- proceeding?

1.4 Thesis Outline (0.5 p)

1.5 To place in chapters above

This master thesis is motivated by a use case from practical work at the Berlin court of audit (Rechnungshof von Berlin; RHvB). The auditors often are faced with the problem that they need information that is provided as natural language or in tables inside of unstructured documents, i.e. in PDF files. The goal of this thesis is benchmarking methods for automated information extraction from specific tables from PDF files.

Ideally, the data extraction pipeline is able to autonomously * identify the pages with the tables of interest. * identify the tables of interest on these pages. * extract the information as provided into a structured table

(e.g. as JSON, a csv file or HTML code). * transform the data into a given schema, stripping all aggregated values.

It should extract the values without errors. It would be nice if the computation time and energy consumption is as low as possible.

A more realistic approach, that is also beneficial to satisfy the AI Act (keine Entscheidung ohne menschliche Beteiligung), is an assistant system, that helps extracting information. Key features to get the human into the loop already at the step of information extraction for such an assistant might be:

- showing the results together with the systems confidence.
- showing the results next to the values of the source.
- allowing in place adjustments to the extracted data.

A sound decision making is only possible if the information the decision is based on is valid.

1.6 RHvB

- what does the RHvB do
- why is this important
- what does it not do yet (because data source is missing)

1.7 Datenverfügbarkeit

- keine Regelung, in welcher Form der Rechungshof die Daten, die er benötigt, bereitgestellt zu bekommen hat

Das Gesetz zur Förderung der elektronischen Verwaltung (EGovG) wurde erlassen, “um die Verwaltung effektiver, bürgerfreundlicher und effizienter zu gestalten.” (BMI, Referat O2, 2013)

§ 12 EGovG

- Vorhaben zur Datenkatalogisierung innerhalb der Verwaltung angestoßen, aber noch nicht richtig gestartet
- Vornehmlich für Bürger*innen Zugang

1.8 Unstrukturierte Daten

- Beispielbilder

1.8.1 Portable Document Format

- print optimized
- Table structure information gets lost
- Bild und Textextract

Chapter 2

Literature review (less than 10 p)

(5 to 10 lines)

- overview of subchapters
- relevance for reader (Gutachter)
- link to previous chapter
- relevant basic tasks

2.1 Basic terms

2.2 Technological topic (related work)

- most important papers
- connection of papers (timeline)
- what used, what not?
- extending existing paper?

2.3 optinal more topics like previous

2.4 Summary (0.5 p)

- lessons learned
- link to goal thesis
- link to next chapter

2.5 To place in chapters above

2.6 Table extraction tasks

2.6.1 Difficulties

- Beispielbilder

2.7 Document Extrtaction Process

2.7.1 Document Layout Analysis

An important step in the process of extracting information from documents is to recognize the layout of a document (Zhong et al., 2019).

Getting the order of texts correct align captions to tables and figure identify headings, tables and figures

One of the most popular datasets used for training and benchmarking is PubLayNet (see PubLayNet on paperswithcode.com). It contains over 360_000 document automatically annotated images from scientific articles publicly available on PubMed Central (Zhong et al., 2019, p. 1). This was possible, because the articles have been provided in PDF and XML format. For the annotations most text categories (e.g. text, caption, footnote) have been aggregated into one category. <– is this a problem for later approaches where a visual and textual model work hand in hand to identify e.g. table captions?

Manual annotated datasets often were limited to several hundred pages. Deep learning methods need a much larger training dataset. Previously optical character recognition (OCR) methods were used.

Identify potentially interesting pages with text / regex search. Check if there is a table present on this page.

Object detection

2.7.1.1 Vision Grid Transformer

2.7.2

2.8 Tools

2.8.1 TableFormer

SynthTabNet <– has it: - nested / hierarchical tables, where rows add up to another row? - identifying units and unit cols/rows

Chapter 3

Implementation 8max 5p)

3.1 Speedup with vLLM and batching

3.2 Setup (Dockerfile and PV)

Chapter 4

Methods

4.1 Page identification

The first task to solve, for a fully autonomous solution, is to identify the pages where the tables of interest are located. For benchmarking 74 annual reports from 7 companies have been used. For this benchmark we limit the tables of interest to those that show **Aktiva**, **Passiva** and **Gewinn- und Verlustrechnung**.

In those documents there are 252 pages of interest holding 265 relevant tables. On 13 pages there have been two tables (**Aktiva** and **Passiva**) on a single page. 21 tables are spread over two pages. In 8 documents there have been multiple tables per type of interest, distributed among the three types of tables as following:

type	count
Aktiva	7
GuV	8
Passiva	7

As baselines a simple regex approach as well as a fully sophisticated visual LLM approach have been used.

4.1.1 Baselines

4.1.1.1 Regex based

results potentially dependend on package used for text extraction (Auer et al., 2024, p. 2 f.)

- PyMuPDF
- pypdf
- docling-parse
- pypdfium
- pdfminer.six

pdfminer informs that some pdfs should not be extracted based on their authors will (meta data field)

results dependend on regex pattern

start with pypdf backend and simple regex developed more sophisticated regex based on missed pages

took wrong identified pages as base for a table detection benchmark and n-shot base for llm classification (contrasts)

some tables can't be found without previous ocr; some pages hold image of table and machine readable text

4.1.1.1.1 LLM based

4.1.1.1.2 VLLM based was not implemented

4.2 Table detection

4.2.1 LLM

- table: yes/no
- akiva: yes/no
- multiclass

4.2.2 Vision Model

Yolo

4.3 Information extraction

4.3.1 Baselines

4.3.2 Simple pipeline

- extract text (if document can't be passed directly)
- query LLM directly

4.3.3 Sophisticated approaches

- with pipelines
- Nougat
- maker
- Azure
- docling

Chapter 5

Results

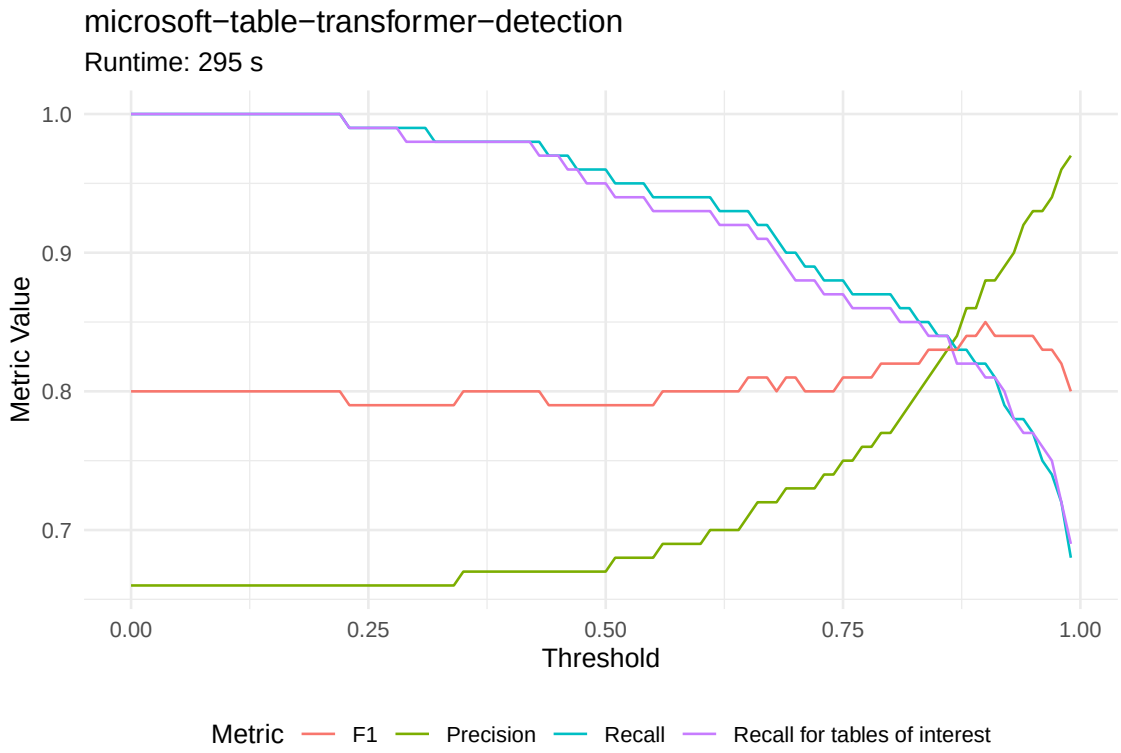
5.1 Page identification

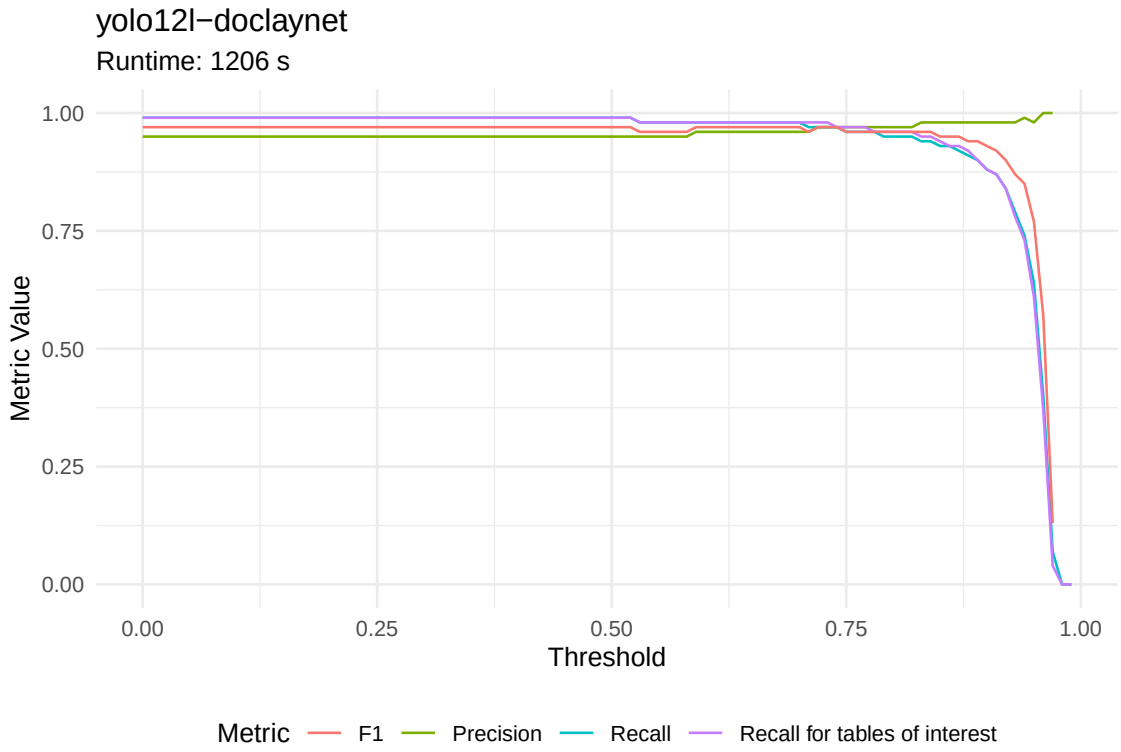
5.1.1 Baselines

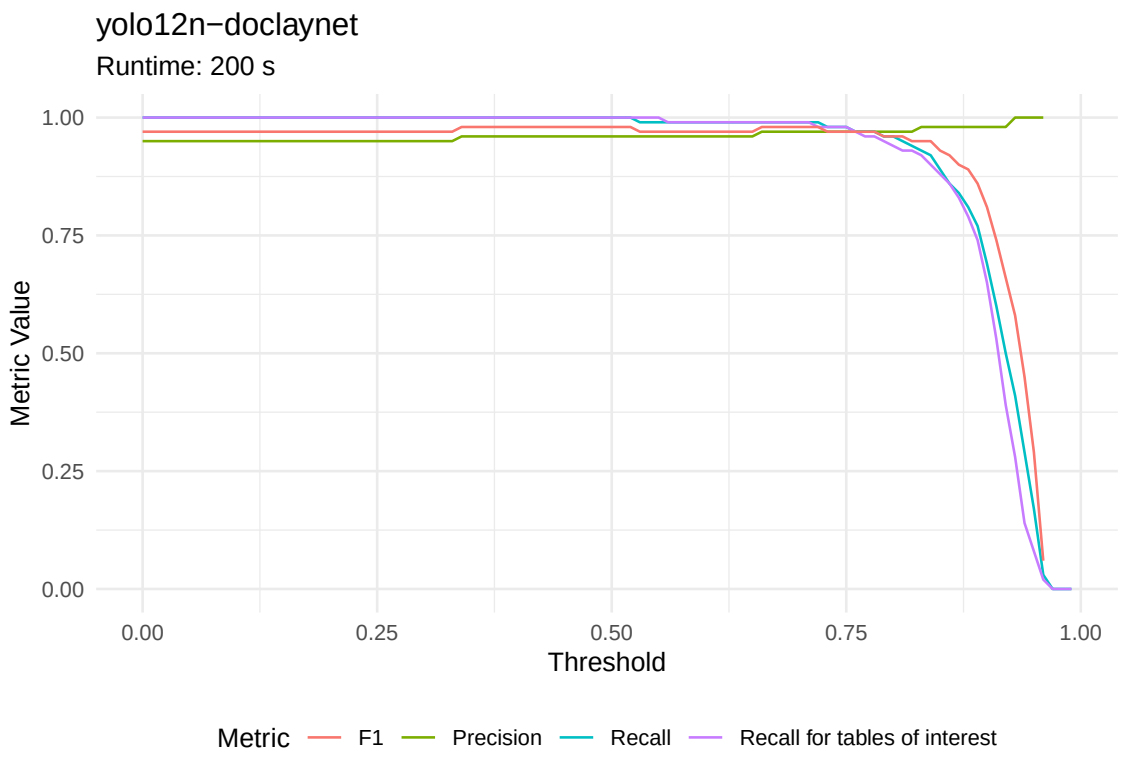
5.1.1.1 Regex based

package	method	precision	recall	F1	runtime_in_s
docling	simple regex	0.24	0.78	0.37	1518.22
pdfium	simple regex	0.24	0.76	0.36	12.42
pdfminer	simple regex	0.24	0.76	0.36	807.15
pdfplumber	simple regex	0.24	0.77	0.37	635.60
pymupdf	simple regex	0.24	0.78	0.37	23.48
pypdf	simple regex	0.25	0.77	0.38	277.23
pdfium	exhaustive regex restricted	0.19	0.92	0.31	24.29
docling	exhaustive regex	0.14	0.93	0.24	1476.34
pdfium	exhaustive regex	0.14	0.93	0.24	14.60
pdfminer	exhaustive regex	0.14	0.91	0.24	797.80
pdfplumber	exhaustive regex	0.14	0.91	0.24	627.78
pymupdf	exhaustive regex	0.14	0.94	0.24	22.32
pypdf	exhaustive regex	0.14	0.93	0.24	256.05

5.2 Table detection






yolo12n−doclaynet
Runtime: 200 s
Metric Value
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Threshold
Metric
F1
Precision
Recall
Recall for tables of interest

llm	parameters	method	loop	F1_Aktiva	runtime_in_s
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	law_context	0	0.2173913	238.30
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	law_context	1	0.2078580	238.17
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	law_context	2	0.2111959	238.66
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	law_context	3	0.2120051	238.52
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	law_context	4	0.2138365	238.47
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	rag_examples	0	0.2187500	289.46
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	rag_examples	1	0.2164276	280.12
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	rag_examples	2	0.2034346	285.80
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	rag_examples	3	0.2018349	343.86
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	rag_examples	4	0.2127108	321.34
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	random_examples	0	0.2114650	211.59
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	random_examples	1	0.2109375	211.00
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	random_examples	2	0.2167742	220.06
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	random_examples	3	0.2190352	214.12
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	random_examples	4	0.2182285	216.64
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	top_n_rag_examples	0	0.2210797	226.02
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	top_n_rag_examples	1	0.2167742	224.53
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	top_n_rag_examples	2	0.2210663	225.72
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	top_n_rag_examples	3	0.2185515	271.77
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	top_n_rag_examples	4	0.2113402	235.59
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	zero_shot	0	0.2180851	122.94
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	zero_shot	1	0.2145695	123.13
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	zero_shot	2	0.2198675	123.00
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	zero_shot	3	0.2207622	123.20
deepseek-ai_DeepSeek-R1-Distill-Qwen-32B	32B	zero_shot	4	0.2198391	123.08
google_gemma-3-27b-it	NA	law_context	0	0.7407407	455.92
google_gemma-3-27b-it	NA	law_context	1	0.7441860	625.80
google_gemma-3-27b-it	NA	law_context	2	0.7476636	626.03
google_gemma-3-27b-it	NA	law_context	3	0.7465438	625.90
google_gemma-3-27b-it	NA	law_context	4	0.7383178	626.01
google_gemma-3-27b-it	NA	rag_examples	0	0.8000000	3827.93
google_gemma-3-27b-it	NA	rag_examples	1	0.7960199	3908.06
google_gemma-3-27b-it	NA	rag_examples	2	0.7980296	3896.02
google_gemma-3-27b-it	NA	rag_examples	3	0.7960199	3895.44
google_gemma-3-27b-it	NA	rag_examples	4	0.7920792	3880.57
google_gemma-3-27b-it	NA	random_examples	0	0.8526316	2653.36
google_gemma-3-27b-it	NA	random_examples	1	0.8404255	2921.28
google_gemma-3-27b-it	NA	random_examples	2	0.8350515	2920.58
google_gemma-3-27b-it	NA	random_examples	3	0.8350515	2933.60
google_gemma-3-27b-it	NA	random_examples	4	0.8125000	2930.62
google_gemma-3-27b-it	NA	top_n_rag_examples	0	0.7902439	1436.34
google_gemma-3-27b-it	NA	top_n_rag_examples	1	0.7980296	1443.84
google_gemma-3-27b-it	NA	top_n_rag_examples	2	0.7941176	1435.03
google_gemma-3-27b-it	NA	top_n_rag_examples	3	0.7941176	1434.89
google_gemma-3-27b-it	NA	top_n_rag_examples	4	0.7941176	1434.64
google_gemma-3-27b-it	NA	zero_shot	0	0.6952790	134.25
google_gemma-3-27b-it	NA	zero_shot	1	0.6952790	206.05
google_gemma-3-27b-it	NA	zero_shot	2	0.6986900	206.14
google_gemma-3-27b-it	NA	zero_shot	3	0.6956522	206.08
google_gemma-3-27b-it	NA	zero_shot	4	0.6982759	206.11
google_gemma-3-4b-it	NA	law_context	0	0.6153846	317.37
google_gemma-3-4b-it	NA	law_context	1	0.6293103	421.29
google_gemma-3-4b-it	NA	law_context	2	0.6127660	421.03
google_gemma-3-4b-it	NA	law_context	3	0.6075949	421.67
google_gemma-3-4b-it	NA	law_context	4	0.6127660	421.01
google_gemma-3-4b-it	NA	rag_examples	0	0.5882353	2445.47
google_gemma-3-4b-it	NA	rag_examples	1	0.5779817	2546.31
google_gemma-3-4b-it	NA	rag_examples	2	0.5777778	2537.95
google_gemma-3-4b-it	NA	rag_examples	3	0.5909091	2548.47
google_gemma-3-4b-it	NA	rag_examples	4	0.5765766	2529.68
google_gemma-3-4b-it	NA	random_examples	0	0.7263158	1705.17
google_gemma-3-4b-it	NA	random_examples	1	0.6868687	1863.38

Chapter 6

Discussion

Chapter 7

Conclusion

References

Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Ramis, C. B., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., Lima, R. T. de, Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., & Staar, P. W. J. (2024). *Docling Technical Report.* arXiv. https://doi.org/10.48550/arXiv.2408.09869

BMI, Referat O2 (Ed.). (2013). *Minikommentar zum Gesetz zur Förderung der elektroni- schen Verwaltung sowie zur änderung weiterer Vor- schriften.*

Zhong, X., Tang, J., & Yepes, A. J. (2019). *PubLayNet: Largest dataset ever for document layout analysis.* arXiv. https://doi.org/10.48550/arXiv.1908.07836

Appendix A

Web UI