BHT 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.

This can be seen as a reimplementation study of "Extracting Financial Data from Unstructured Sources: Leveraging Large Language Models" - a paper published by Li et al. (2023). The key differences are the application on German documents using open source large language models.

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.

Reading advices

The author recommends to read the thesis in its digital gitbook version instead of the PDF version. Furthermore, the author recommends to read the thesis (any version) on a screen that is larger than 21" and has at least full HD resolution[1]. The more, the merrier.

😊

Notes

- Qwen 2.5 hat zweiseitige GuV von IBB entdeckt und zur Anpassung der Ground Truth
- Google gemma war mit alter Klassifikation erfolgreich (anderer Prompt, mehr Seiten)

implementation nach methods

- rerun real table analysis after creating more gold standarsd with good working approach
- found mistakes in gold standard with the llm results

[1] Most of the time the thesis was inspected at a third of the authors 42" screen with 4k resolution. For inspecting the large overview graphics it is a very handy tool the author recommends every data scientist or software developer.

Goals and Learnings

Achieved:

- thesis with bookdown
- docker image creation
- cluster orchestration
- llm usage
- guided decoding

Missed:

- Administrating a k8s cluster
- Fine tuning a model
- using small language models
- training a lm
- using vllms

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?

missing law to access digital data and no law to choose the format of the data extensible Business Reporting Language as a standard changing from HGB to IFSR

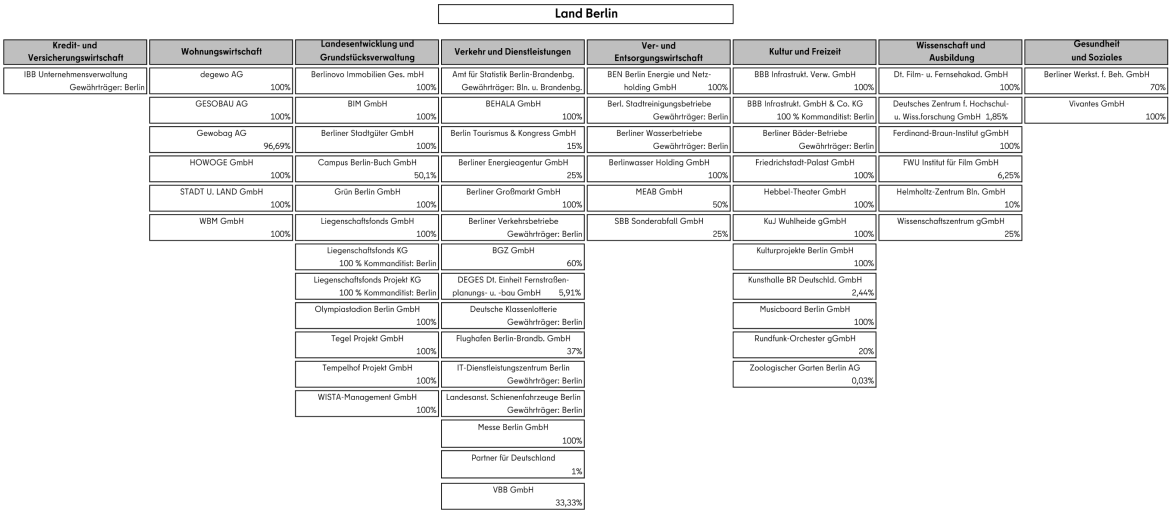


Figure 1.1: Companies Berlin has holds share at

1.2 Objectives

The sixth division at RHvB is auditing the companies Berlin is a stakeholder of. Basic information they have to process are the balance sheets and profit and loss accounting. Those information is provided via their

annual reports in form of PDF files. The provided annual reports often differ from the publicly available ones in matter of information granularity and design and are treated as non public information. Automate the extraction of those information would be a good starting point for AI assisted information retrieval from PDFs for the RHvB overall.

It is important to get numeric values totally accurate; numeric values are difficult to handle for langauge models

- 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?

Research questions and hypothesss

Q1: Can a LLM (large language model) be used to efficiently extract financial information from German annual reports? Q2. Can LLMs be used to identify the page of interest automatically?

Q3: Can confidence scores be used to head up the human in the loop on which results to double check? (How can sources of the automatic extraction being communicated down stream in order to make double checking easy before making decisions?) Q4: Can contextual information from similar documents reduce errors made during table extraction? Q5: What are characteristics of financial tables that make it hard for LLMs to identify / extract them? (How does the length and complexity of financial documents (e.g., multi-column layouts, nested tables) affect table extraction performance?)

1.3 Methodology (1 p)

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

Experimental / Comparative Research • Reimplementing framework(s) • Comparing / Benchmarking • Frameworks • Models • Methods • Use cases • Ablation test

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
- parameter vs active parameter

2.1 NLP history

2.2 Basic terms

2.3 Supervised Learning Approaches

2.3.1 Generalized Linear Models

2.3.2 Random Forest

XGBoost not used finally, because calculation SHAP (SHapley Additive exPlanations) values for XGBoost model took to long for just a first glimpse on what might influence the extraction.

2.3.3 Large Language Models

2.3.3.1 Embeddings

2.3.3.2 Neural networks in NLP

2.3.3.3 Attention / Multi-Head

2.3.3.4 Transformers

2.3.3.5 Encoder

2.3.3.6 Decoder

2.3.3.7 BERT

2.3.3.8 Bi-Encoder

2.3.3.9 Mixture of Experts

2.3.3.10 Guided decoding

generation template strict (closed) vs open

2.3.3.11 Classification trained models (not used)

Soft max

2.3.3.12 Few-shot Learning

2.3.3.13 RAG

2.3.3.14 GPT (Generative Pretrained Transformers)

2.3.4 Information extraction

closed-domain vs open-domain

2.4 Data balancing

2.4.1 Under sampling, oversampling

2.5 Evaluation Metrics

2.5.1 For classification

2.5.2 For regression

2.6 Technological topic (related work)

- LLM generation

- structured output
- Fewshot
- context length can be harmful
- most important papers
- connection of papers (timeline)
- what used, what not?
- extending existing paper?

2.7 Term frequency

2.7.1 Extraction of numeric values

99.5 % or 96 % accuracy for extracting financial data from Annual Comprehensive Financial Reports (Li et al., 2023) In the untabulated test, GPT-4 achieved an average accuracy rate of 96.8%, and Claude 2 achieved 93.7%. Gemini had the lowest accuracy rate at 69%. (ebd.)

Too many hallucinated values when it was NA instead (Grandini et al., 2020)

2.8 optinal more topics like previous

2.9 Summary (0.5 p)

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

2.10 To place in chapters above

2.11 Table extraction tasks

2.11.1 Difficulties

- Beispielbilder

2.12 Document Extrtaction Process

2.12.1 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

2.12.1.1 Vision Grid Transformer

2.12.2

2.13 Tools

2.13.1 TableFormer

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

Chapter 3

Methods

norm gpu hours

3.1 Data

- companies Beteiligungsbericht
- number found Jahresberichte
- number used Jahresberichte first rows
- number used Jahresberichte Aktiva Tabellen

3.2 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 a baseline a simple regex approach was used.

3.2.1 Baselines

3.2.1.1 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

3.2.1.1.1 LLM based

3.2.1.2 Term frequency based

3.2.1.2.1 VLLM based was not implemented

3.3 Table detection

Can be used to narrow down set of possible pages

Can be used to focus only on the table content (measure if correct area was identified would be necessary)

Vision model as baseline

3.3.1 LLM

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

3.3.2 Vision Model

Yolo

3.3.3 Docling and Co

3.3.3.0.1 VLLM based was not implemented

3.4 Information extraction

3.4.1 Baselines

simple regex?

3.4.2 Simple pipeline

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

3.4.3 Sophisticated approaches

not implemented

- with pipelines
- Nougat
- maker
- Azure
- docling

Chapter 4

Implementation (max 5p)

4.1 Speedup with vLLM and batching

4.2 Setup (Dockerfile and PV)

Chapter 5

Results

5.1 Page identification

As described in A.2.1 open source libraries have been used to extract the text from the annual reports.

5.1.1 Baseline: Regex

Building a sound regular expression often is an iterative process. In a first approach a very simple one was implemented.

Comparing the differences in the metrics based on the different text extraction libraries it can be said that the extracted text is very similar but not identical. Since the resukts are not depending on the used text extraction library the *exhaustive regex restricted* has only been run with the fast text extraction library *pdfium*. The results of the regex based page identification are presented in the following tables.

- look into details where they differ and if it is because of a line break or whitespace ?

Due to the imbalanced distribution of the classes the accuracy is not a good metric to compare the performance of the different methods. The number of pages of interest is much smaller than the number of irrelevant pages. Therefore, precision, recall and F1 score are presented as well.

The regular expressions can be found in the appendix (see 5.1.1).

General bad precision. Increasing recall degrades precision even further. number of pages positive identified total; used as subset for table identification task

Table 5.1: Comparing page identification metrics for different regular expressions for classification task 'Aktiva'

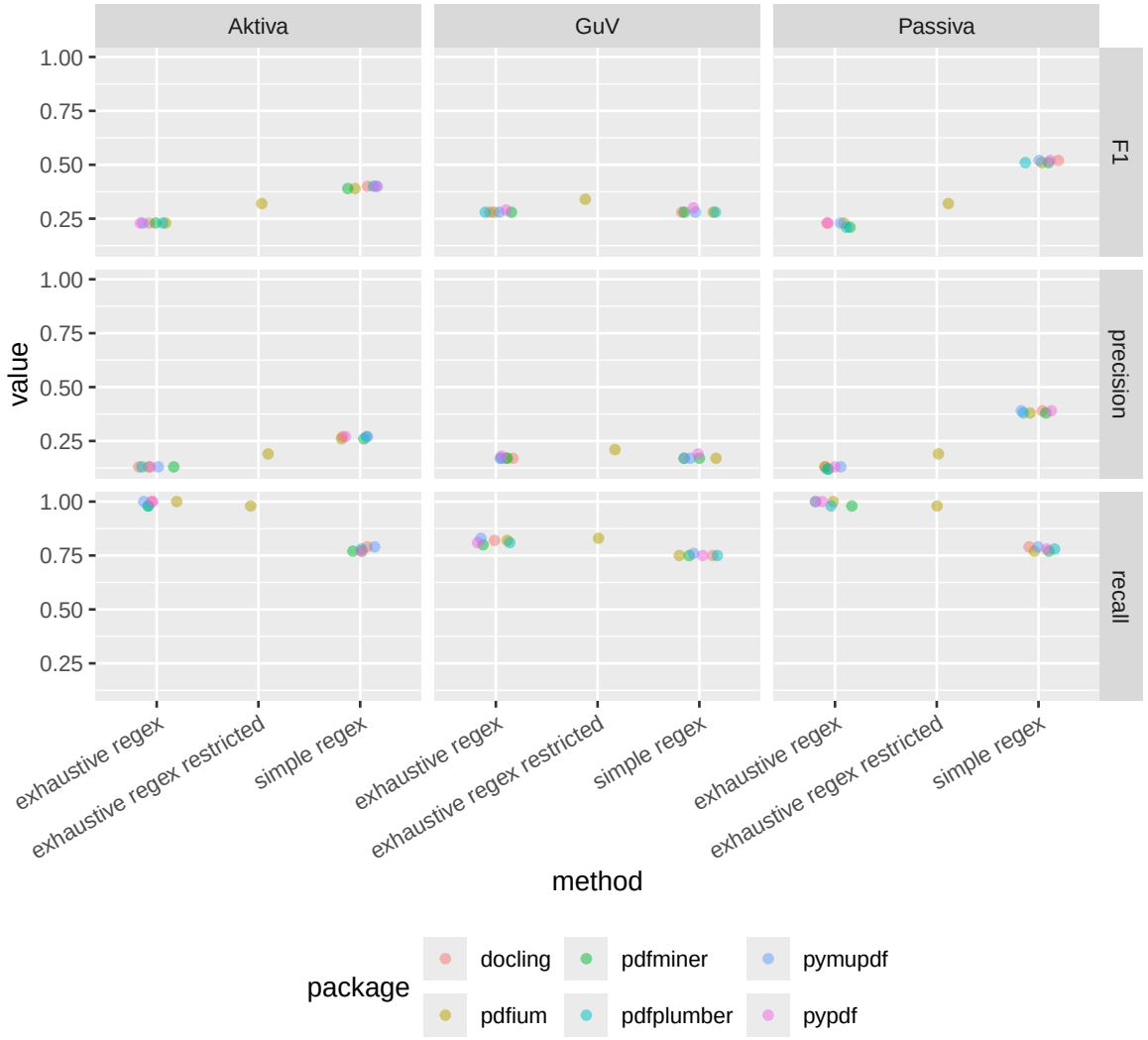
method	stat	precision	recall	F1
simple regex	mean	{0.267}	0.778	{0.397}
simple regex	sd	0.005	0.01	0.005
exhaustive regex restricted	mean	0.19	0.98	0.32
exhaustive regex restricted	sd	NA	NA	NA
exhaustive regex	mean	0.13	{0.993}	0.23
exhaustive regex	sd	0	0.01	0

Table 5.2: Comparing page identification metrics for different regular expressions for classification task 'Passiva'

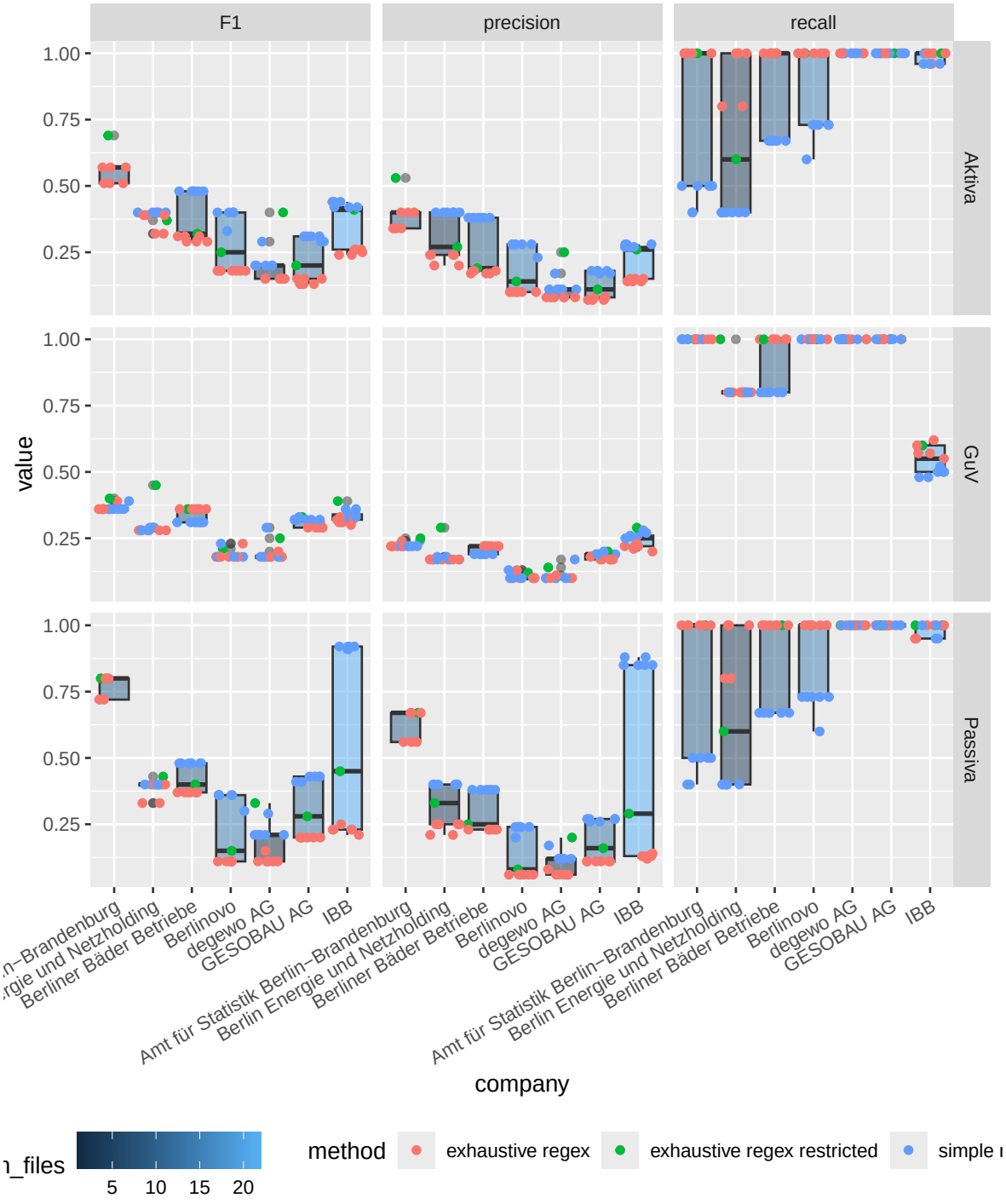
method	stat	precision	recall	F1
simple regex	mean	{0.385}	0.78	{0.515}
simple regex	sd	0.005	0.009	0.005
exhaustive regex restricted	mean	0.19	0.98	0.32
exhaustive regex restricted	sd	NA	NA	NA
exhaustive regex	mean	0.127	{0.993}	0.223
exhaustive regex	sd	0.005	0.01	0.01

Table 5.3: Comparing page identification metrics for different regular expressions for classification task 'Gewinn und Verlustrechnung'

method	stat	precision	recall	F1
simple regex	mean	0.173	0.752	0.283
simple regex	sd	0.008	0.004	0.008
exhaustive regex restricted	mean	{0.21}	{0.83}	{0.34}
exhaustive regex restricted	sd	NA	NA	NA
exhaustive regex	mean	0.172	0.815	0.282
exhaustive regex	sd	0.004	0.01	0.004



Results by company?



5.1.2 Table of Contents understanding

An optional step for larger documents in Li et al. (2023) framework is to identify the pages of interest based on the table of contents (TOC). This would be more efficent than processing the whole document with an LLM. The TOC in a PDF can be given explicit and machine readable or it can be presented in form of text on any page. Of course it can be missing completely as well.

- For a lot of short annual reports one can find the tables of interest within the first eight pages as well.
- calculate and add Qwen, Gemini or LLama results? <– No time!

5.1.2.1 Text based

Li et al. (2023) used the table of contents to identify the pages of interest. In their approach the table of contents is extracted from the text. Based on their observation, that the TOC that "ACFRs typically spans no more than the initial 165 lines of the converted document" (p. 20), they use the first 200 lines of text.

My expectation was to find the TOC within the first five pages. Often we find way less than 200 lines of text on the five first pages (see Figure 5.1). Some files are not machine readable without OCR and thus show zero lines in the first five pages as well.

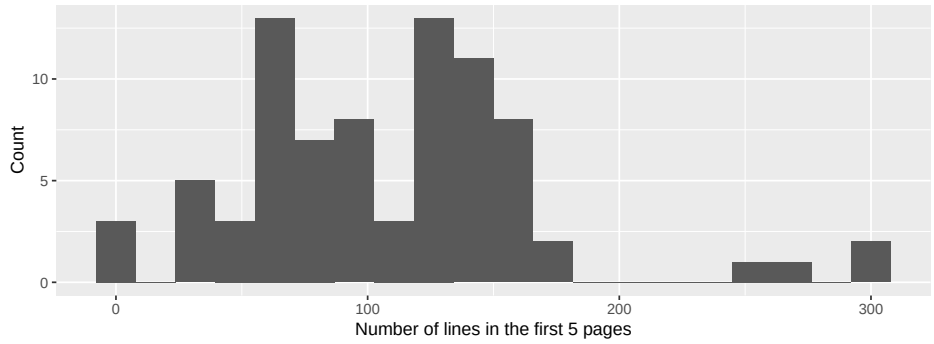


Figure 5.1: Histogram of the number of lines in the first 5 pages of the annual reports

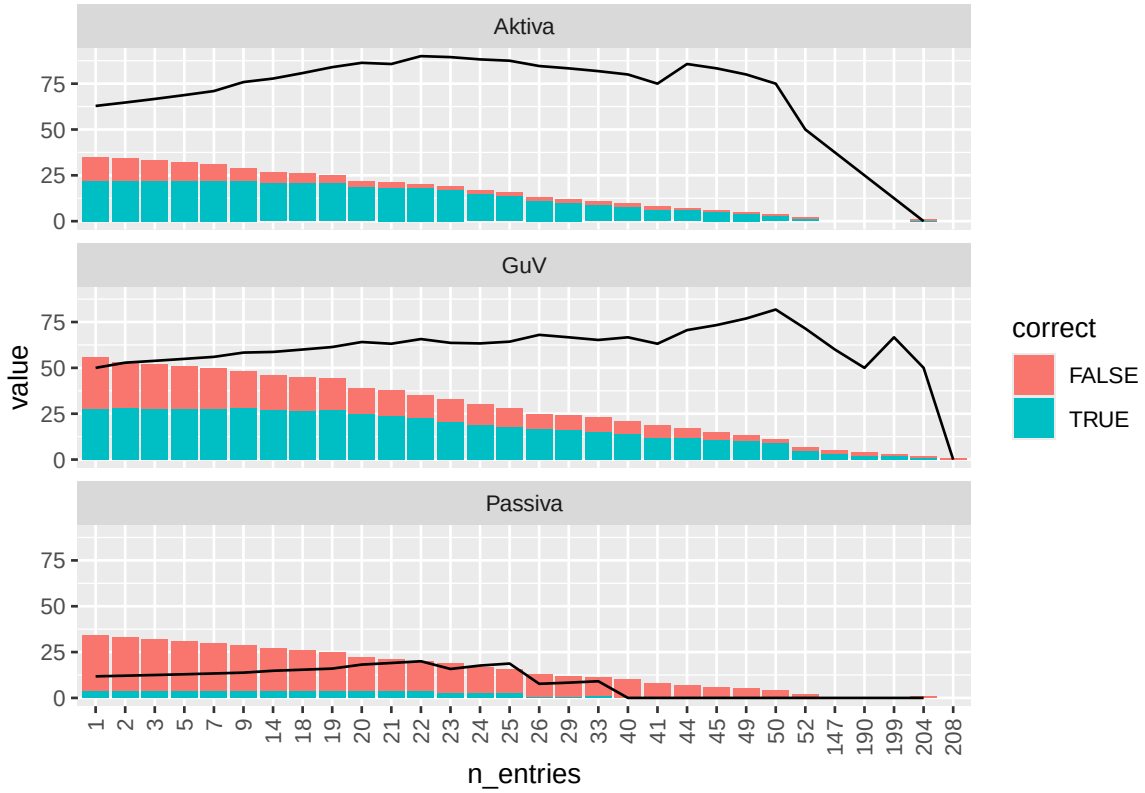
5.1.2.1.1 First five pages A request to Mistral results in 63 strings that should represent a table of contents among the first five pages [strings not checked in detail].

5.1.2.1.2 First 200 lines A request to Mistral results in 68 strings that should represent a table of contents among the first five pages [strings not checked in detail].

5.1.2.1.3 Machine readable TOC based To limit the text and hopefully increase the quality of the input data one can work with the TOC representation embedded within the PDF files. From 80 annual reports 43 files do have a machine readable separate table of contents and 37 do not have one.

One can see that correct predictions for the page range are more probable when the TOC has a medium number of entries. It is possible to drop PDFs with less than 9 without loosing a single correct prediction. This means that for PDFs with TOC with less then 9 entieres the LLM was not able to make a correct prediction. This is not surprising since neither *Bilanz* nor *Gewinn- und Verlustrechnung* are mentioned there.

Almost no influence if TOC is passed formated as markdown or json. With the json formated TOC it found two more correct page ranges (single test run). It was testes because the relation *page_number* heading and value might have been clearer in json for a linear working LLM.



5.1.2.2 Comparison of the different approaches

- toc analysis
- cleaned measures

The LLM performed best on the machine readable TOC. It resulted in highest ratio of correct page ranges as well as in highest absolute numbers even though there were least available TOC.

Values can be higher than 80, the total number of PDF files, since there can be multiple tables of interested for the same type in a single document or a table of interest can span two pages. Since the prompt for the LLM was not elaborated enough to cover cases, where there are multiple tables of interest for a single type that are not placed on concurrent pages, one could argue to drop those files from the analysis. This does not change the results significantly, since there are only few files with more than one table of interest per type.

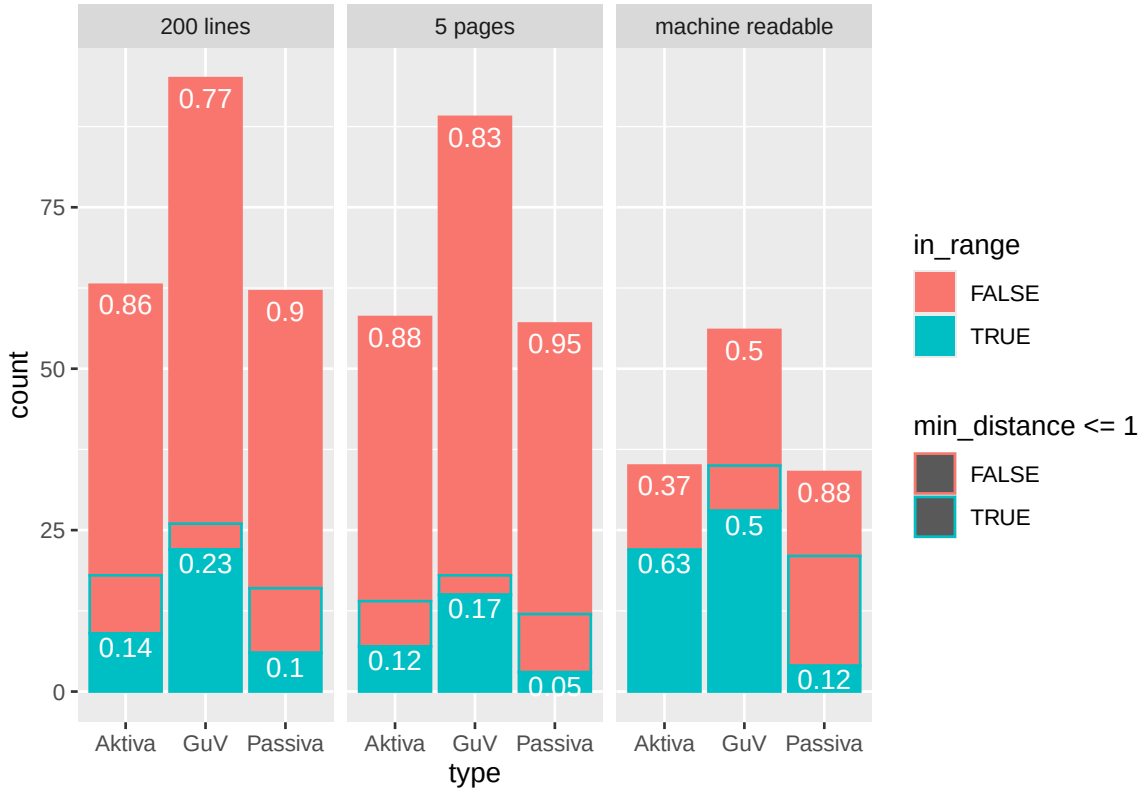
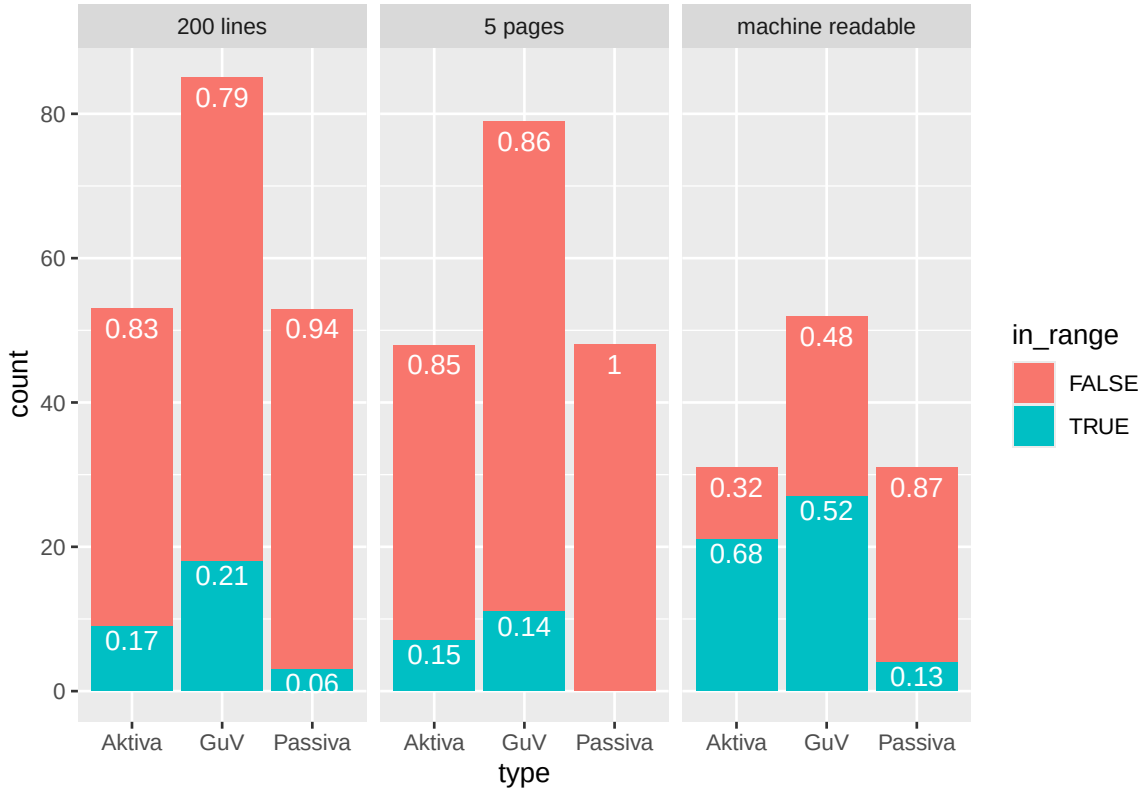


Figure 5.2: Comparing number of fount TOC and amount of correct and incorrect predicted page ranges

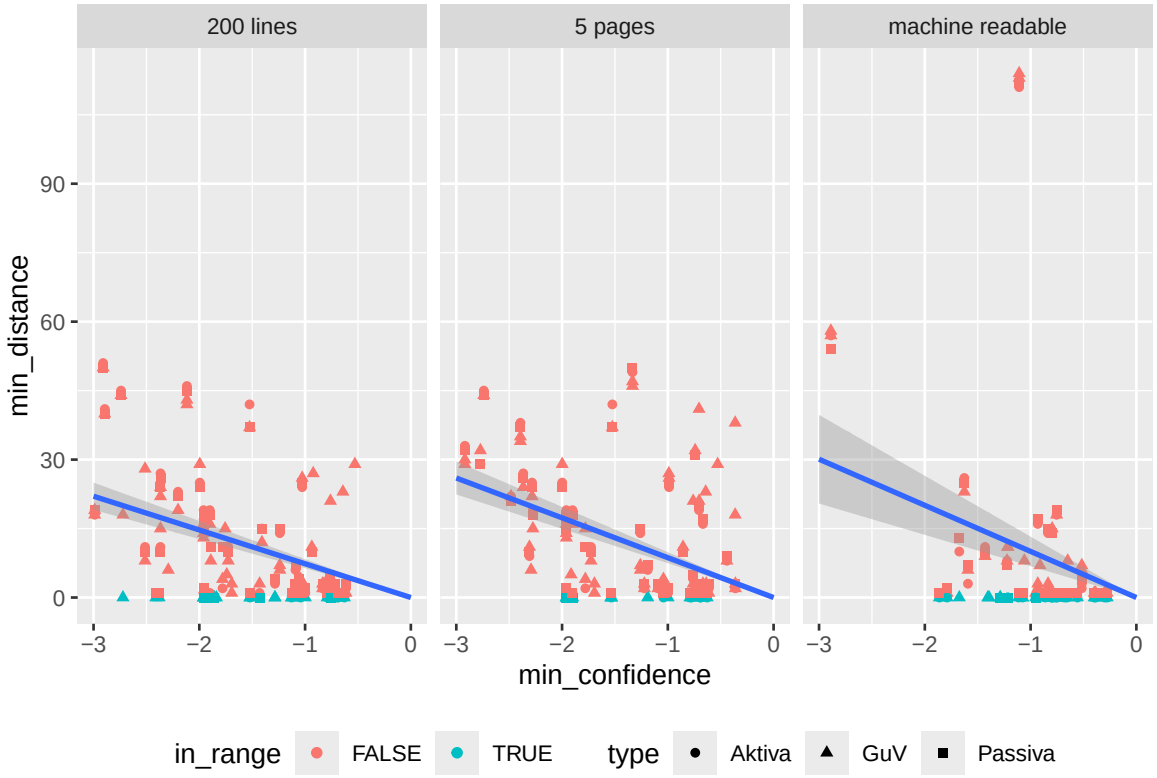
benchmark_type	mean range	SD range
5 pages	{2.35}	{1.79}
200 lines	2.78	2.33
machine readable	2.81	4.35



Besides a single group that was predicted far off for the machine readable TOC approach the LLM reported higher confidence for the correct page ranges and got the ranges less far off. But it did not predict the smallest ranges.

Table 5.4: Comparing GPU time for page range prediction and table of contents extraction

Benchmark Type	Page range predicting	TOC extracting
5 pages	{0.56}	{2.19}
200 lines	0.57	3.8
machine readable	0.63	NA



In general the LLM performed worst to identify the correct page range for *Passiva*. The median distance is one page bigger than for *Aktiva* and *Gewinn- und Verlustrechnung*. This makes sense for *Aktiva* because the *Passiva* is often on the next page but the predicted page range for *Aktiva* and *Passiva* are often identical. Furthermore the predicted page range for *Aktiva* is often only a single page wide. Thus the *Passiva* on the next page is not inside the predicted page range.

This problem was solved by explicitly mentioning that assets and liabilities are both part of the balance sheets for the five pages and 200 lines approach but not for the machine readable TOC one.

A pragmatic way would be to use the machine readable TOC approachs prediction for the *Aktiva* page range and add one to the end page to get the *Passiva* page range. Beside the problem to predict a correct page range for *Passiva* the machine readable TOC approach was very effective and is also pretty efficient if one counts in the effort the LLM driven TOC extraction takes.

5.1.3 Classification with LLMs

structured outputs forcing to answer with a *yes* or *no* for binary task or with *Aktiva*, *Passiva*, *GuV* or *other* for multi classification task

top n accuracy

out of company vs in compnay rag

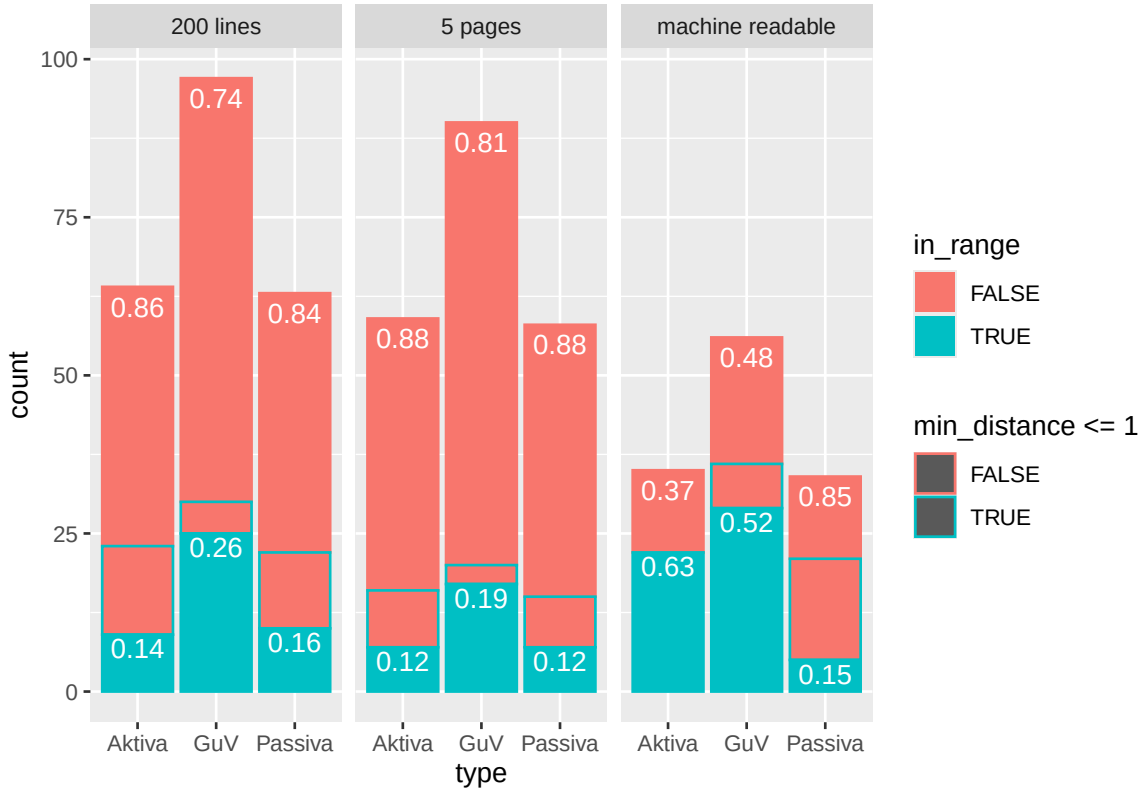


Figure 5.3: Comparing number of fount TOC and amount of correct and incorrect predicted page ranges

model_family	model	classification_type	method_family	n_examples	f1_s
mistralai	mistralai_Ministral8BInstruct2410	GuV	n_rag_examples	3	0.93
meta-llama	metallama_Llama4Scout17B16EInstruct	GuV	n_rag_examples	3	0.92
mistralai	mistralai_Ministral8BInstruct2410	Passiva	n_rag_examples	3	0.92
mistralai	mistralai_Ministral8BInstruct2410	Aktiva	n_rag_examples	3	0.92
Qwen	Qwen_Qwen2.532BInstruct	GuV	n_rag_examples	1	0.87
meta-llama	metallama_Llama4Scout17B16EInstruct	Passiva	n_rag_examples	3	0.85
Qwen	Qwen_Qwen2.532BInstruct	Aktiva	n_rag_examples	1	0.84
Qwen	Qwen_Qwen3235BA22BInstruct2507	Aktiva	n_rag_examples	3	0.84
meta-llama	metallama_Llama4Scout17B16EInstruct	Aktiva	n_rag_examples	3	0.83
Qwen	Qwen_Qwen2.532BInstruct	Passiva	n_rag_examples	1	0.79
microsoft	microsoft_phi4	Aktiva	law_context	1	0.68
microsoft	microsoft_phi4	Passiva	law_context	1	0.65
google	google_gemma327bit091	Passiva	n_rag_examples	1	0.54
google	google_gemma327bit091	Aktiva	n_rag_examples	1	0.51
tiiuae	tiiuae_Falcon310BInstruct	Passiva	n_random_examples	1	0.5
google	google_gemma327bit091	GuV	n_rag_examples	1	0.49
tiiuae	tiiuae_Falcon310BInstruct	Aktiva	n_rag_examples	1	0.44
tiiuae	tiiuae_Falcon310BInstruct	GuV	top_n_rag_examples	1	0.33

5.1.3.1 Binary classification

Could be more efficient to predict “is any of interest” and then which type, because dataset is highly imbalanced.

25 models from 6 haven been benchmarked among 5 methods

Most models have been used up to 3 examples for the context.

The best combination of model and method for each method family is presented in the following table. It is clear that the Google Gemma models are performing worst.[1] Surprisingly Mistral 2410 is the best performing model for all three prediction tasks even though it only has 8B parameters.

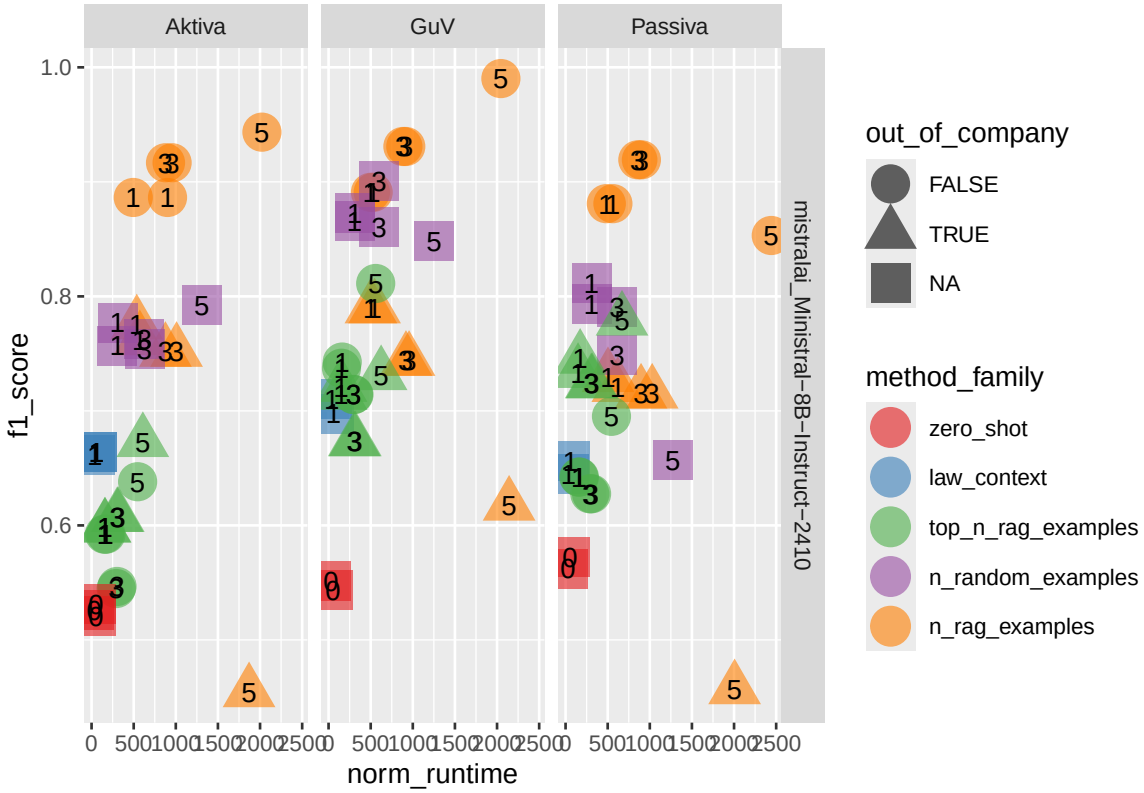
It is interesting that the predictions do not get better by providing more and more examples. Especially for the n-rag-example approach we find a significant drop in the F1 score if the examples pages come from different companies annual reports. This is caused by a sever recall drop. But also for the n-ranom-example approach we see this for the prediction of class Passiva.

Recall better with examples from same company. Precision better without.

We can also see that the prediction performance is stable.[2]

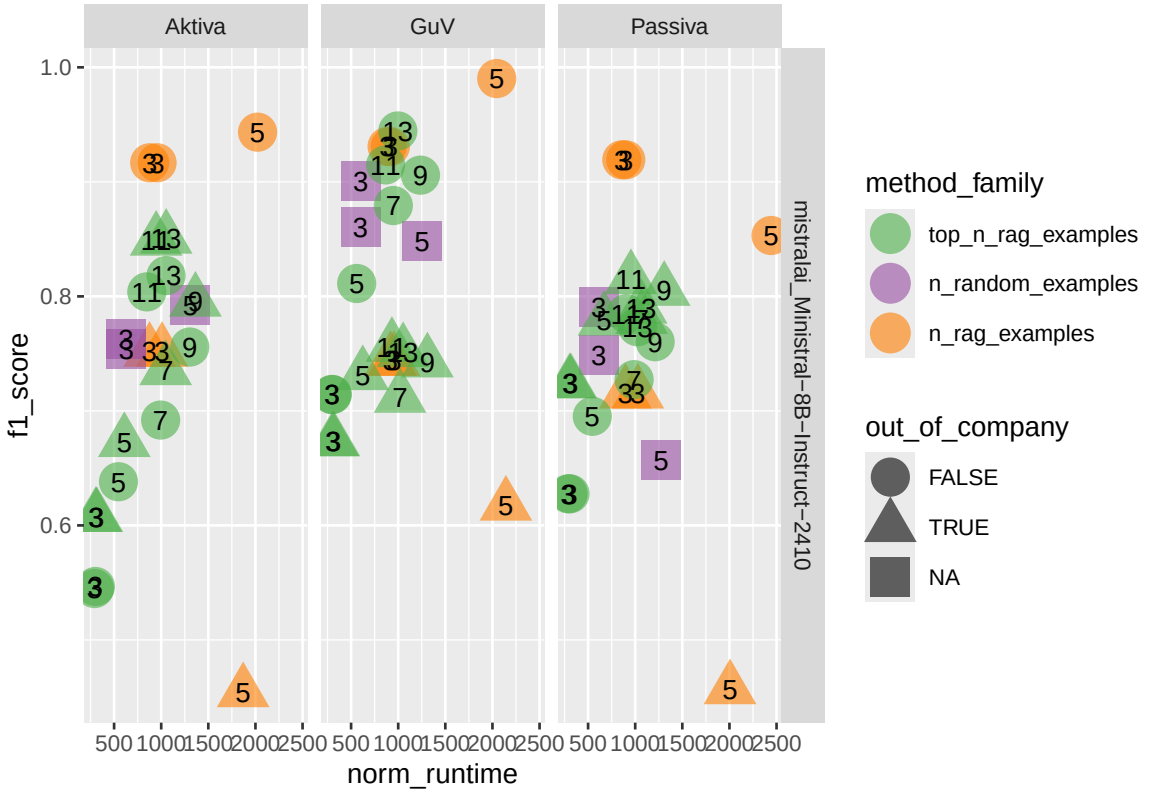
[1] This is not due to a temporary technical problems caused by a bug in the transformers version shipped with the vllm 0-9-2 image. Those problems have been overcome. The performance stays bad.

[2] Earlier experiments on a subset of the pages have been run five times indicating stable results. Running the experiments up to tree times in this very task indicate this as well.



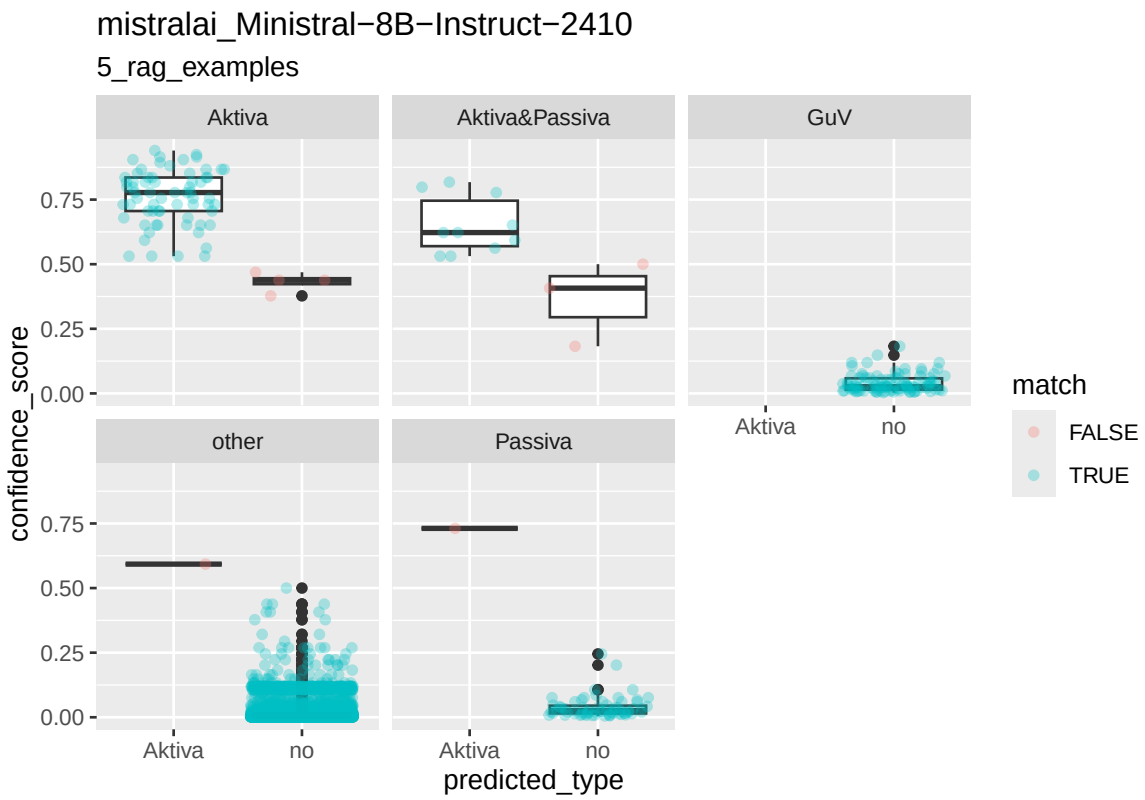
- f1
- multiple models
- best model detail (different methods / settings)

The experiments for the best performing model, Ministral-8B-Instruct-2410, have been extended by methods with even more examples. Especially for the top-n-rag-example approach to get a better comparable picture based on the real number of examples / context length.

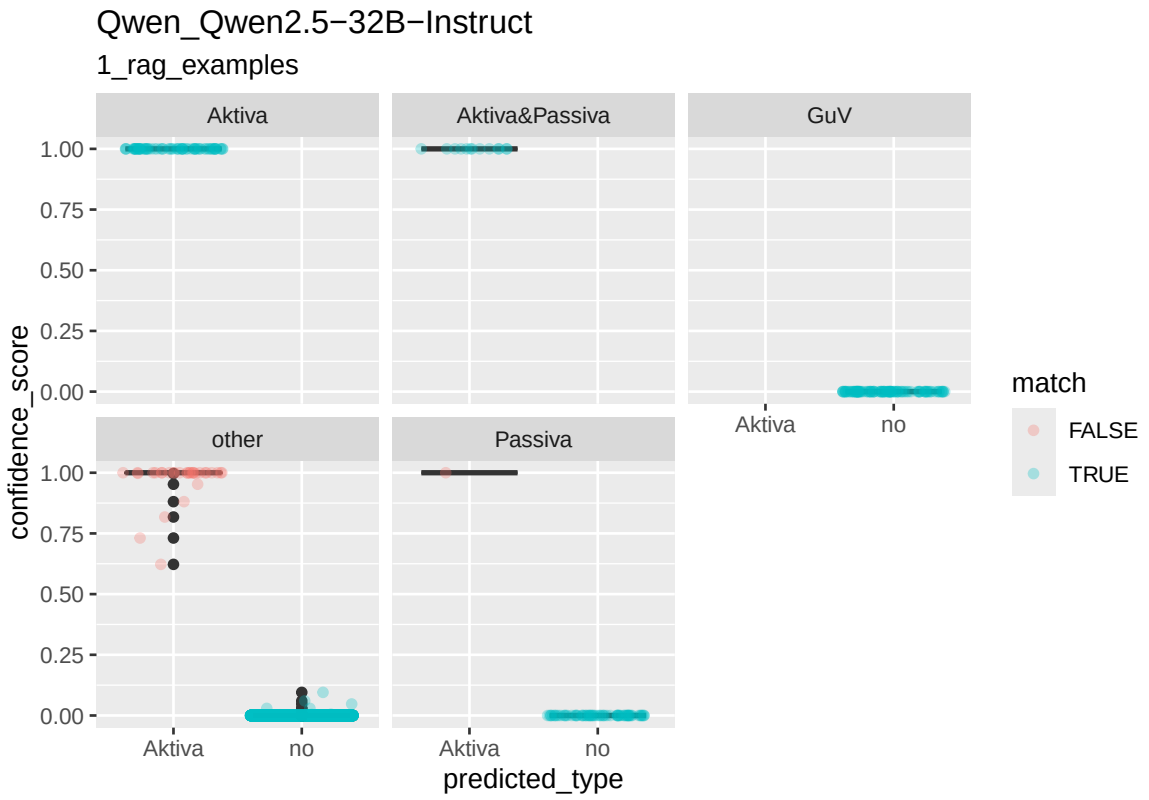


approach	classification	n_example	target	other	sum
n_random_examples	binary	1	1	1	4
n_random_examples	binary	3	3	1	6
n_random_examples	binary	5	5	2	11
n_random_examples	multi	1	1	1	4
n_random_examples	multi	3	3	3	12
n_random_examples	multi	5	5	5	20
n_rag_examples	binary	1	1	1	4
n_rag_examples	binary	3	3	1	6
n_rag_examples	binary	5	5	2	11
n_rag_examples	multi	1	1	1	4
n_rag_examples	multi	3	3	3	12
n_rag_examples	multi	5	5	5	20
top_n_rag_examples	binary	1	NA	NA	1
top_n_rag_examples	binary	3	NA	NA	3
top_n_rag_examples	binary	5	NA	NA	5
top_n_rag_examples	binary	7	NA	NA	7
top_n_rag_examples	binary	9	NA	NA	9
top_n_rag_examples	binary	11	NA	NA	11
top_n_rag_examples	binary	13	NA	NA	13
top_n_rag_examples	multi	1	NA	NA	1
top_n_rag_examples	multi	3	NA	NA	3
top_n_rag_examples	multi	5	NA	NA	5
top_n_rag_examples	multi	7	NA	NA	7
top_n_rag_examples	multi	9	NA	NA	9
top_n_rag_examples	multi	11	NA	NA	11
top_n_rag_examples	multi	13	NA	NA	13

Predictions very accurate. Confidence not always 1. Wrong predictions with often with medium confidence. If Aktiva and Passiva on same page more often Aktiva predicted. Confidence for no displayed as 1-confidence to represent confidence for yes (binary classification).

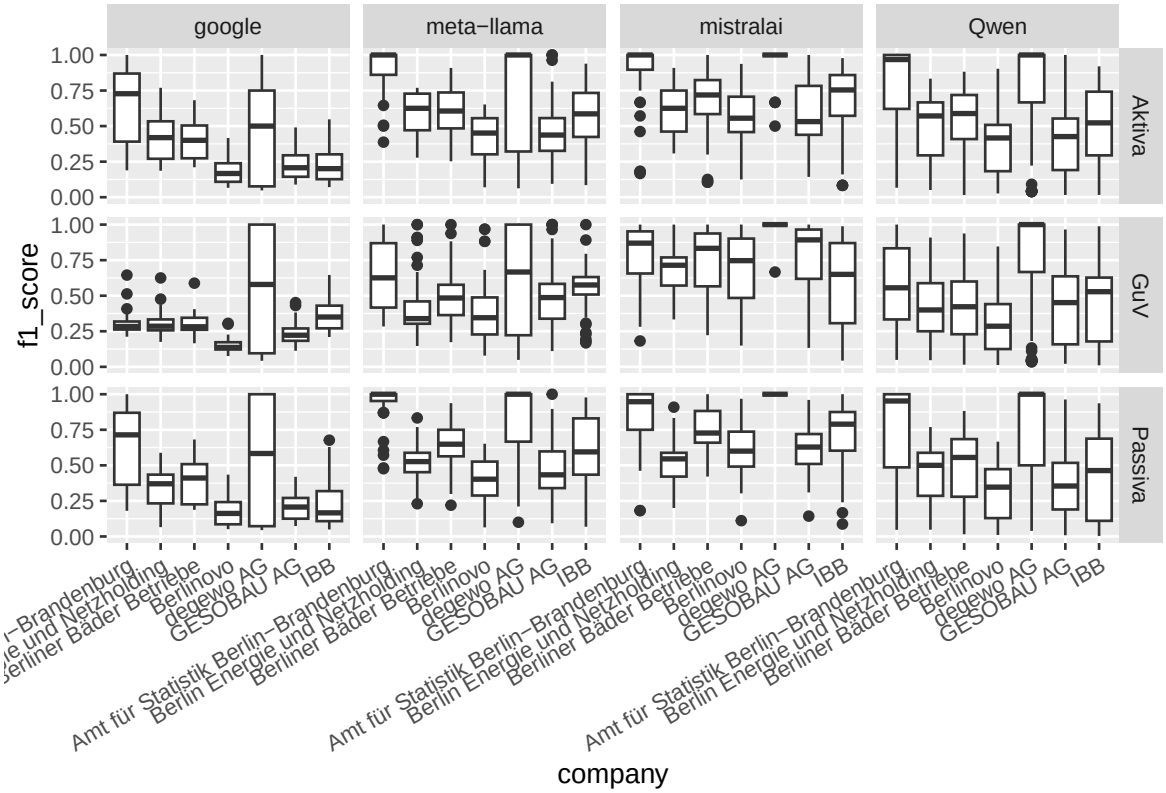


Qwen returns always high confidence even if it is wrong.



- IBB other law
- degewo only one where no ocr is needed

mistral: recall IBB and Netzholding big range meta & mistral: very high precision for Amt für Statistik BBB <- lowest average pagecount (29.3) but IBB has more pages than berlinovo but better precision. No information about which company / report the page is from

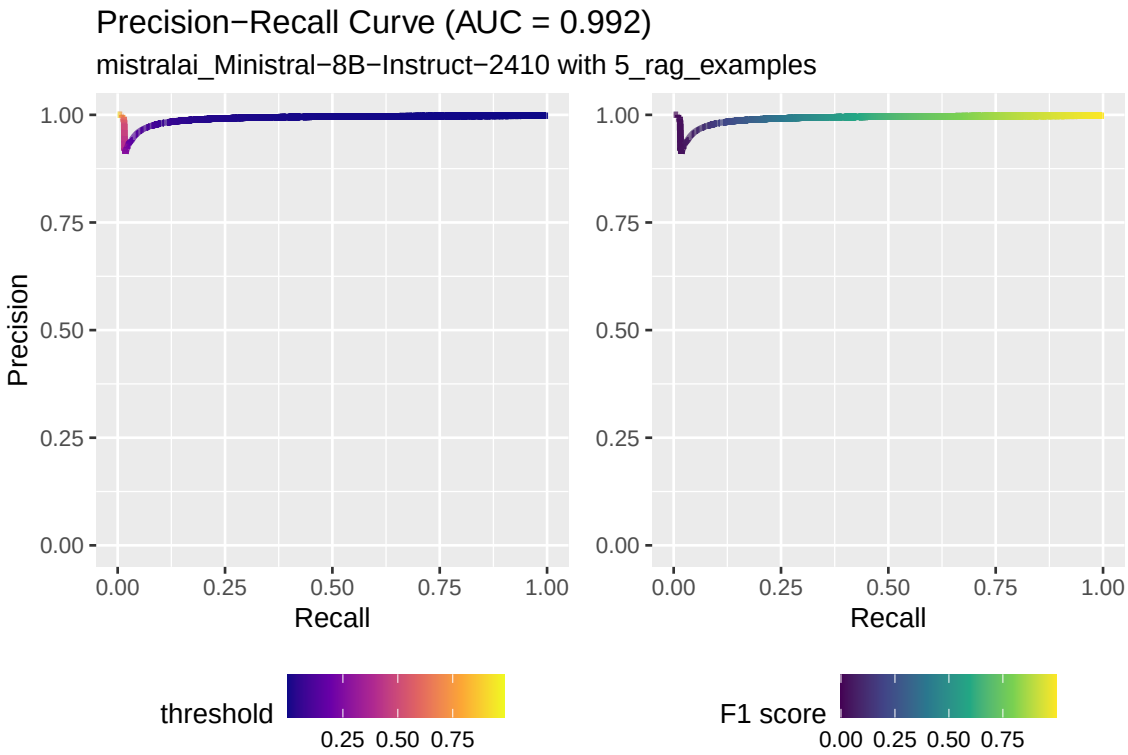


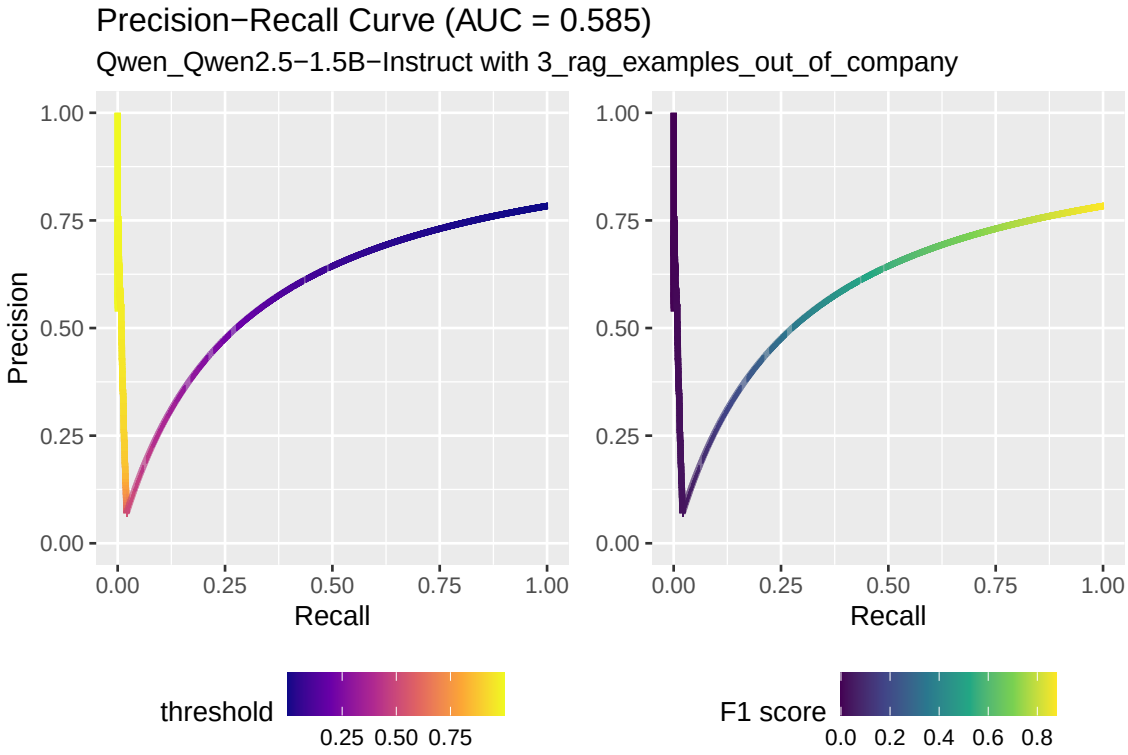
company	n
Amt für Statistik Berlin-Brandenburg	10
Berlin Energie und Netzholding	3
Berliner Bäder Betriebe	10
Berlinovo	15
GESOBAU AG	13
IBB	22
degewo AG	1

company	n
Amt für Statistik Berlin-Brandenburg	10
Berlin Energie und Netzholding	3
Berliner Bäder Betriebe	10
Berlinovo	15
GESOBAU AG	13
IBB	22
degewo AG	7

- Performance makes a jump at a critical parameter number (3B) then slow increase (compare Qwen 2.5)
- Changes unsystematic with new models (see Mistral, Qwen 3 old vs llama 4)

PR curves for all classes look very alike- showing micro average curve


Precision−Recall Curve (AUC = 0.992)
mistralai_Ministral−8B−Instruct−2410 with 5_rag_examples
Precision
Recall
1.00
0.75
0.50
0.25
0.00
threshold
0.25 0.50 0.75
F1 score
0.00 0.25 0.50 0.75


Precision−Recall Curve (AUC = 0.585)
Qwen_Qwen2.5−1.5B−Instruct with 3_rag_examples_out_of_company
Precision
Recall
1.00
0.75
0.50
0.25
0.00
threshold
0.25 0.50 0.75
F1 score
0.0 0.2 0.4 0.6 0.8

model_family	model	metric_type	method_family	n_examples	f1_score	run
meta-llama	metallama_Llama4Scout17B16EInstruct	GuV	n_rag_examples	1	1	254
meta-llama	metallama_Llama4Scout17B16EInstruct	Aktiva	n_rag_examples	3	0.99	445
mistralai	mistralai_MistralLargeInstruct2411	Passiva	n_rag_examples	1	0.99	706
meta-llama	metallama_Llama4Scout17B16EInstruct	Passiva	n_rag_examples	3	0.99	445
mistralai	mistralai_MistralLargeInstruct2411	Aktiva	n_rag_examples	3	0.97	187
Qwen	Qwen_Qwen2.532BInstruct	Aktiva	n_rag_examples	3	0.97	566
Qwen	Qwen_Qwen3235BA22BInstruct2507	GuV	n_rag_examples	3	0.97	111
mistralai	mistralai_MistralLargeInstruct2411	GuV	n_rag_examples	1	0.95	706
mistralai	mistralai_MistralLargeInstruct2411	GuV	n_rag_examples	3	0.95	187
Qwen	Qwen_Qwen2.572BInstruct	Passiva	n_rag_examples	1	0.95	390
google	google_gemma327bit091	Aktiva	n_rag_examples	3	0.88	429
google	google_gemma327bit091	Passiva	n_rag_examples	1	0.81	260
google	google_gemma327bit091	GuV	n_rag_examples	1	0.78	260
tiiuae	tiiuae_Falcon310BInstruct	GuV	n_rag_examples	1	0.7	868
tiiuae	tiiuae_Falcon310BInstruct	Aktiva	n_rag_examples	3	0.69	239
microsoft	microsoft_phi4	Passiva	n_rag_examples	2	0.67	166
microsoft	microsoft_phi4	Aktiva	n_random_examples	1	0.59	493
tiiuae	tiiuae_Falcon310BInstruct	Passiva	top_n_rag_examples	3	0.59	494
microsoft	microsoft_phi4	GuV	n_rag_examples	1	0.45	172

5.1.3.2 Multi classification

bigger models are better with the multi classification task Llama-4-Scout almost perfect F1 for all classes

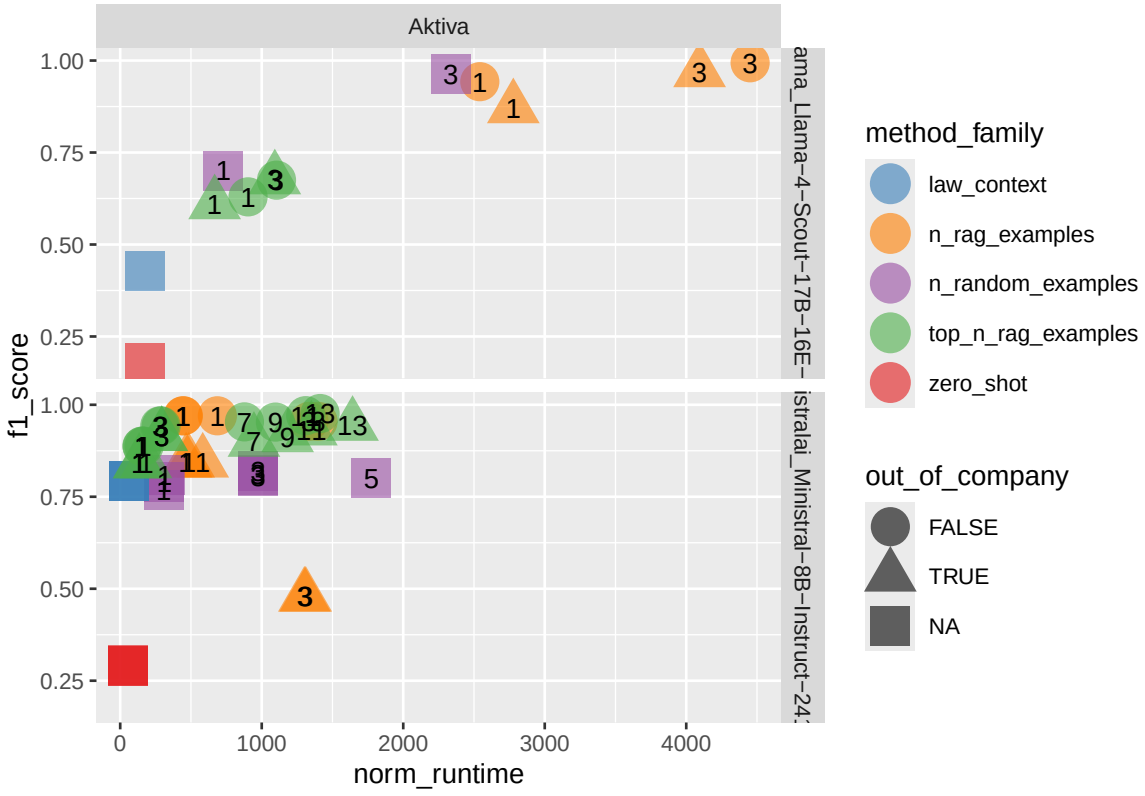
Llama-4-Scout runs fast but needs long to load because it has 109B in total with 17B actives Gemma performs much better than with binary classification

drop with Qwen-14B

Mistral-8B-2410 almost as good as Mistral-123B-2411 but much faster

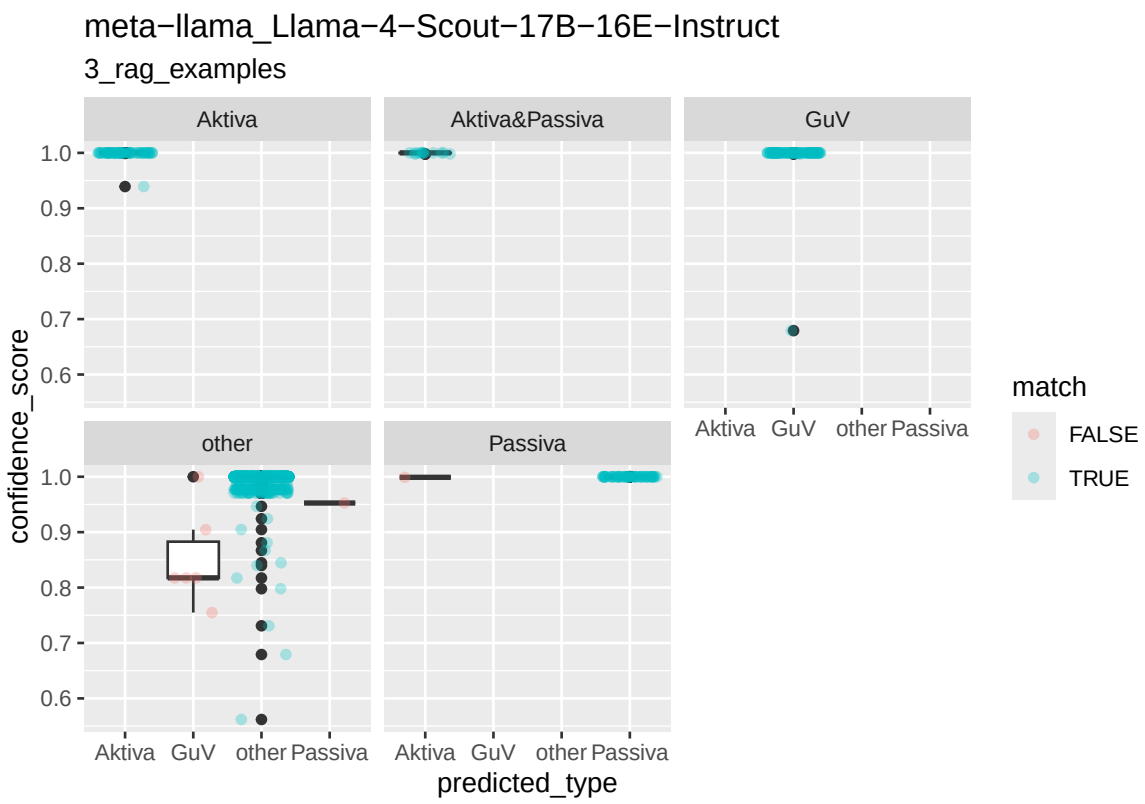
Mistral-2410 reaches good performance already with few examples and can work with law-context approach but more examples don't realy help any further

model_family	model	metric_type	method_family	n_examples	f1_score	runti
mistralai	mistralai_Ministral8BInstruct2410	Aktiva	n_rag_examples	1	0.97	686
mistralai	mistralai_Ministral8BInstruct2410	Passiva	n_rag_examples	1	0.95	686
mistralai	mistralai_Ministral8BInstruct2410	GuV	n_rag_examples	3	0.95	1399
mistralai	mistralai_Ministral8BInstruct2410	GuV	top_n_rag_examples	3	0.95	279
meta-llama	metallama_Llama3.18BInstruct	Passiva	n_rag_examples	1	0.94	593
Qwen	Qwen_Qwen2.53BInstruct	Aktiva	n_rag_examples	1	0.86	492
meta-llama	metallama_Llama3.18BInstruct	Aktiva	top_n_rag_examples	3	0.85	269
google	google_gemma312bit091	Aktiva	n_rag_examples	3	0.84	2733
Qwen	Qwen_Qwen2.53BInstruct	Passiva	law_context	NA	0.81	28
Qwen	Qwen_Qwen2.53BInstruct	GuV	n_rag_examples	1	0.76	492
tiiuae	tiiuae_Falcon310BInstruct	GuV	n_rag_examples	1	0.7	868
tiiuae	tiiuae_Falcon310BInstruct	Aktiva	n_rag_examples	3	0.69	2393
google	google_gemma312bit091	Passiva	n_rag_examples	1	0.68	1259
meta-llama	metallama_Llama3.18BInstruct	GuV	n_rag_examples	1	0.62	593
meta-llama	metallama_Llama3.18BInstruct	GuV	top_n_rag_examples	1	0.62	205
tiiuae	tiiuae_Falcon310BInstruct	Passiva	top_n_rag_examples	3	0.59	494
google	google_gemma312bit091	GuV	top_n_rag_examples	1	0.46	232



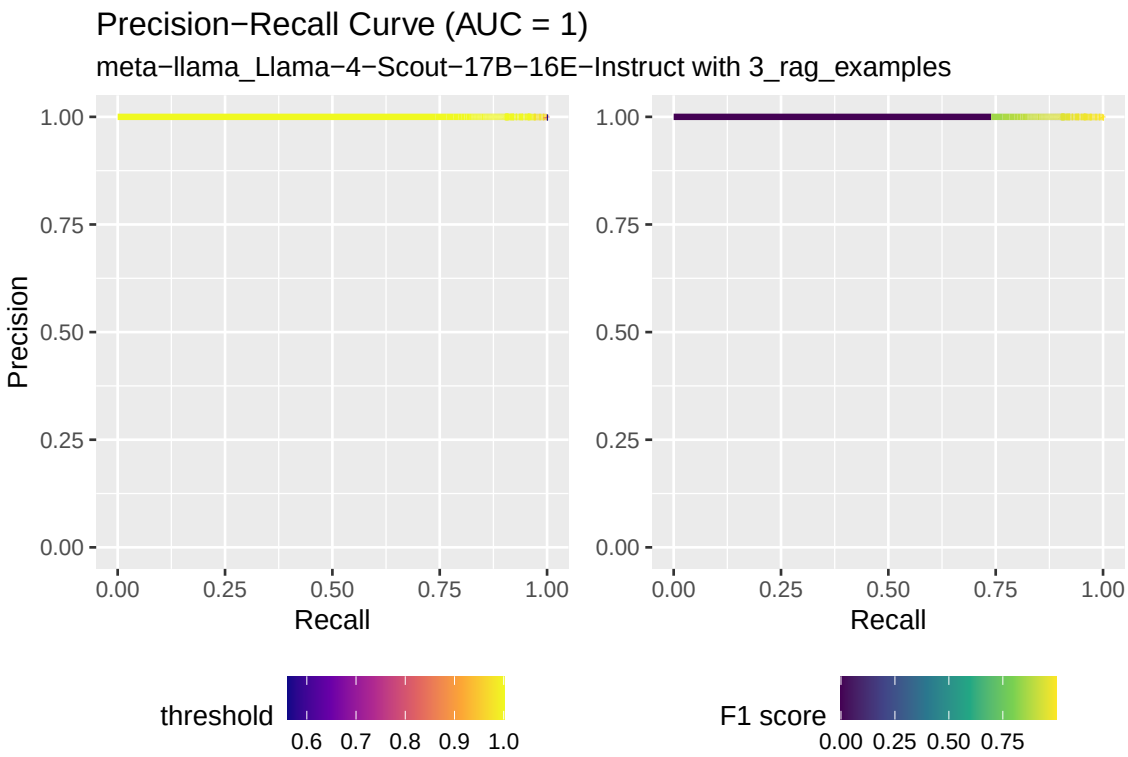
Most of the time pretty confident most problems with class “other” If Aktiva and Passiva on same page it predicts Aktiva. Also one Passiva missclassified as Aktiva No flipped confidence [3]

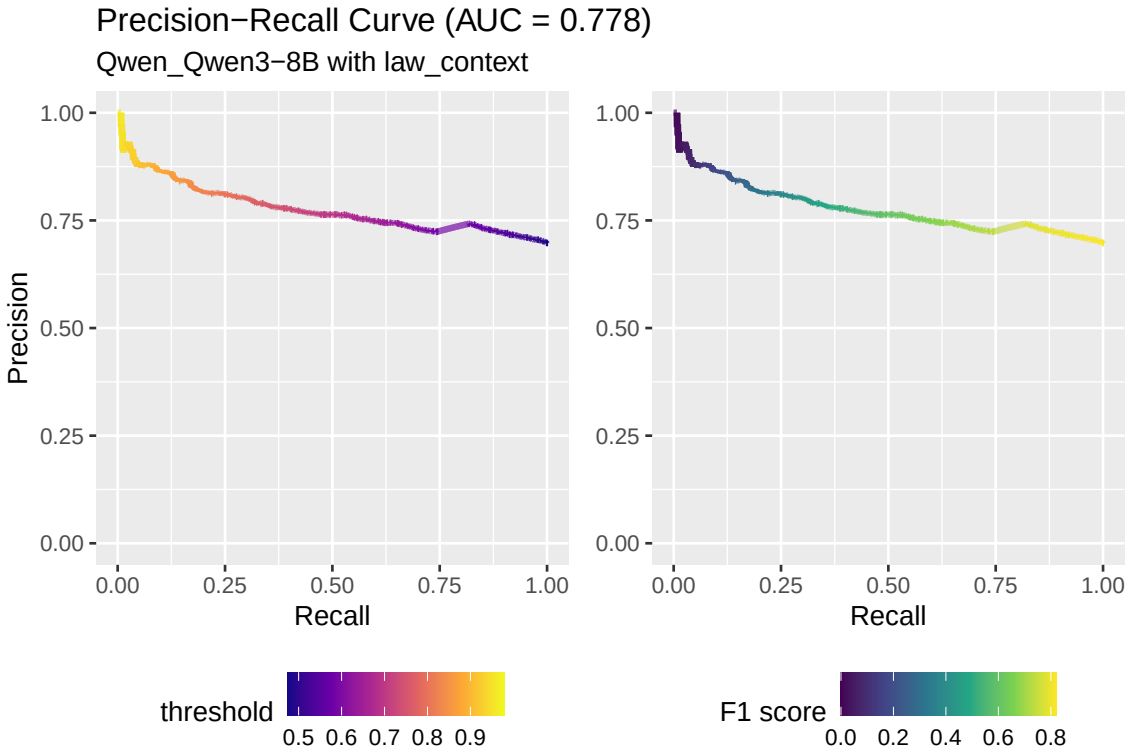
[3] classify framework in needs special models with pooling capability. Would have been interesting but time was limited and would have needed new special models in most cases



Microsoft phi 4 and Falcon 3 only ran with one and two examples because their context window is smaller.

- f1
- multiple models
- best model detail (different methods / settings)


Precision−Recall Curve (AUC = 1)
meta−llama_Llama−4−Scout−17B−16E−Instruct with 3_rag_examples
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
threshold
0.6 0.7 0.8 0.9 1.0
F1 score
0.00 0.25 0.50 0.75


Precision−Recall Curve (AUC = 0.778)
Qwen_Qwen3−8B with law_context
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
threshold
0.5 0.6 0.7 0.8 0.9
F1 score
0.0 0.2 0.4 0.6 0.8

5.1.4 Term frequency based classifier

RandomForest performs much better than a logistic regression Better results with * undersampling * training on all types simultaniousely

5.1.4.1 Two predictors

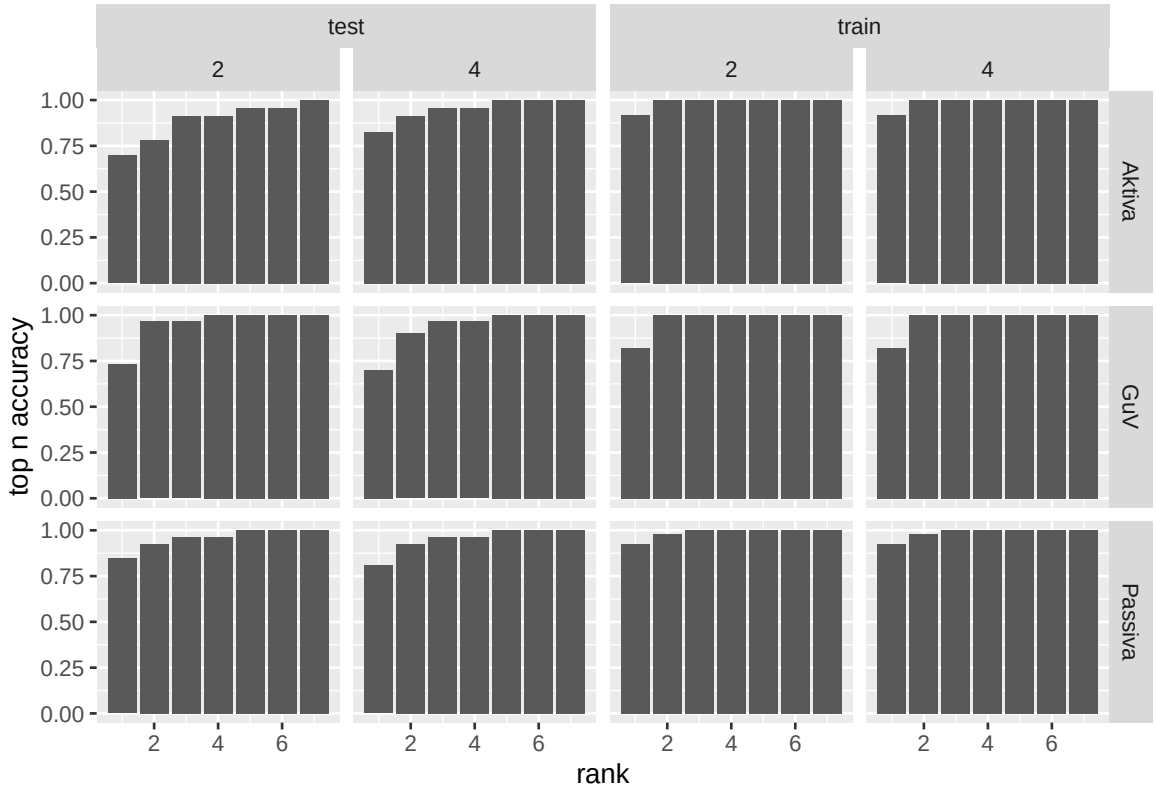
Term frequency of nouns of the law about Aktiva Float freqency (floats divided by word count)

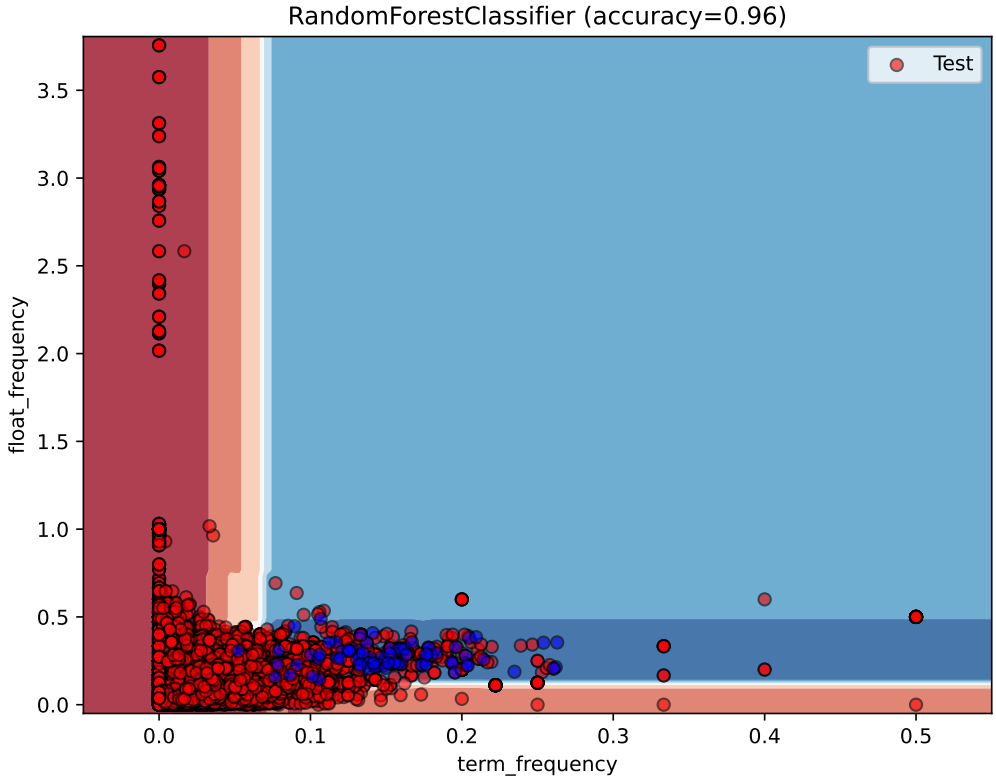
5.1.4.2 Four predictors

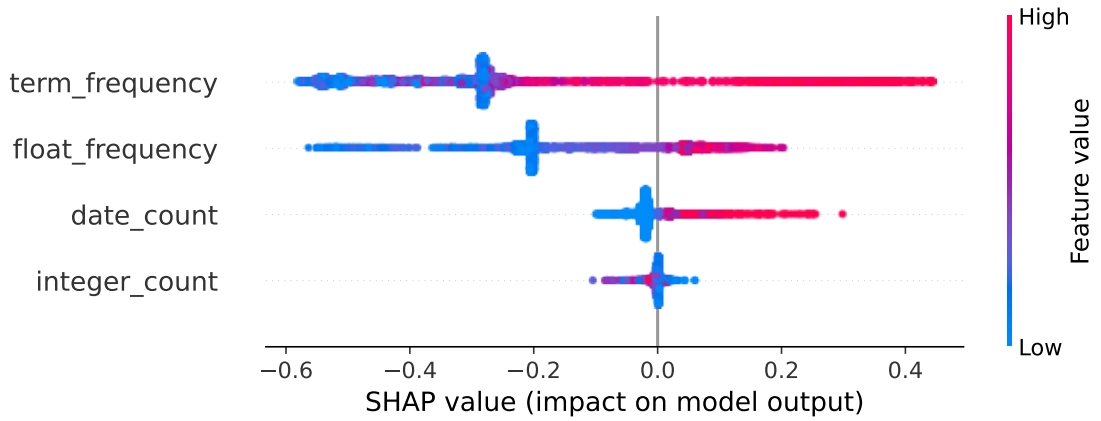
Count of integers Count of dates

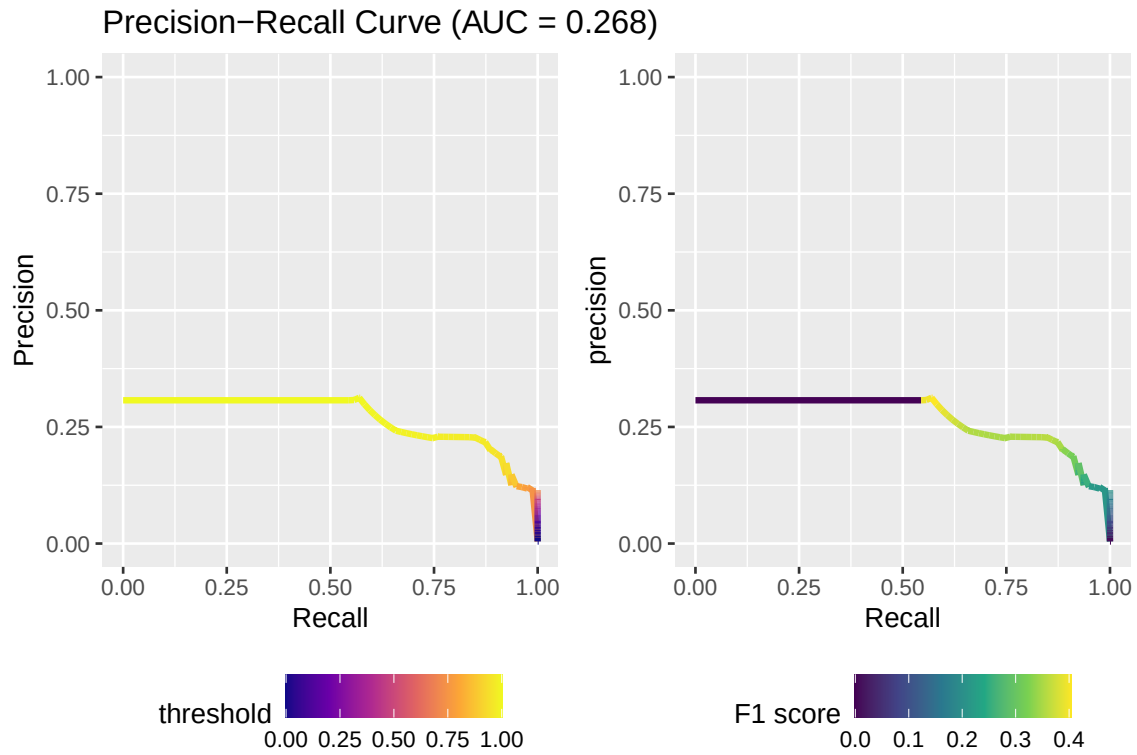
- top 1
- top k

low precision llm linked to position of correct page? numeric frequency?




RandomForestClassifier (accuracy=0.96)
Test
float_frequency
term_frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.1
0.2
0.3
0.4
0.5


term_frequency
float_frequency
date_count
integer_count
−0.6
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Feature value
Low



5.1.5 Comparison

5.1.5.1 Prediction performance

F1 scores for llms are much higher

5.1.5.2 Energy usage and runtime

Multiclassification more effective than three times single classification Combine term frequency with llm approach to limit page range

5.2 Table extraction

5.2.1 Baseline: Regex

The baseline for the table extraction task is set by an approach using regular expressions on the text extract. The approach performs much better[4] on the synthetic dataset compared to the real dataset (see Figure 5.4). Even though, it does not perform perfectly and its performance is more consistent on the text extracted with pymupdf compared to pdfium. Some possible explanations are:

- a duplicated row name[5]

[4] A comparison of the numeric values over all methods can be found in section 5.2.3.

[5] The row *Geleistete Anzahlungen* can be found in two parts of the table and the simple approach just matches the numbers to the first found entry.

- numeric columns extracted separated from row names by extraction libraries
- sums in the same row as the single values[6]
- with pdfium: missing white space[7]
- with pdfium: random line breaks[8]

You can find some examples for incorrect extracted texts in section A.7.

On the real dataset the approach shows a wider spread for the percentage of correct extracted numeric values as well as a considerable number of annual reports where the extraction did not work at all. Interestingly, the used text extraction library has no noticeable influence on the real dataset.

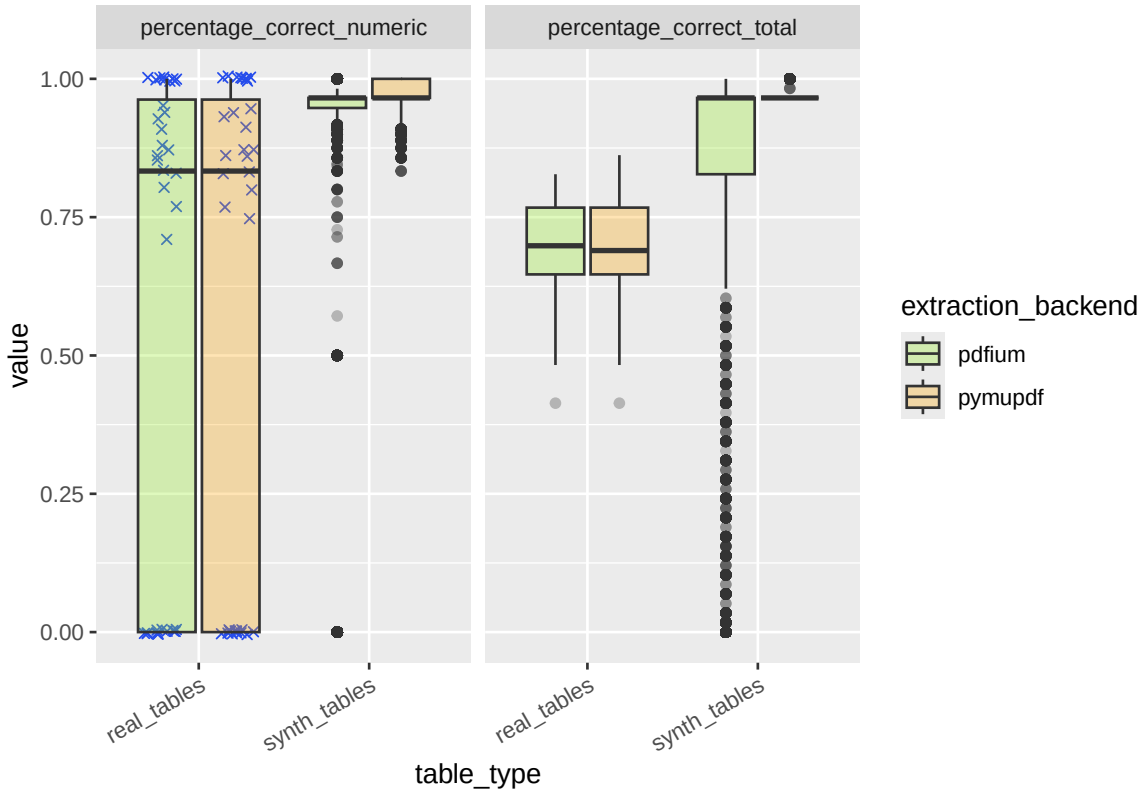


Figure 5.4: Performance overall and on numeric value extraction with regular expressions. Showing single scores for *percentage correct numeric* on real tables to explain wide boxes.

The random line breaks result in some missed row names which is reflected by the bigger spread for NA precision with pdfium on the synthetic dataset (see Figure 5.5). Nevertheless, the NA precision for the majority of the cases is perfect. This is different with the real dataset. The NA precision is found to be at only 0.7.

5.2.1.0.1 Hypotheses The formulated hypotheses have been evaluated visually using the dependence and beeswarm plots from the shapviz library based on the SHAP values calculated with a random forest.

5.2.1.0.1.1 Real dataset There are multiple hypotheses that don't get supported by the visual results (see Figure A.1). The pretty surprising results are:

[6] In this case the regex (regular expression) takes the sum as the value for the previous year.
[7] This can form unexpected numeric patterns or prevent the row names to be recognized.
[8] The approach takes care of line breaks between words, but not within. This leads to unrecognized row names as well.

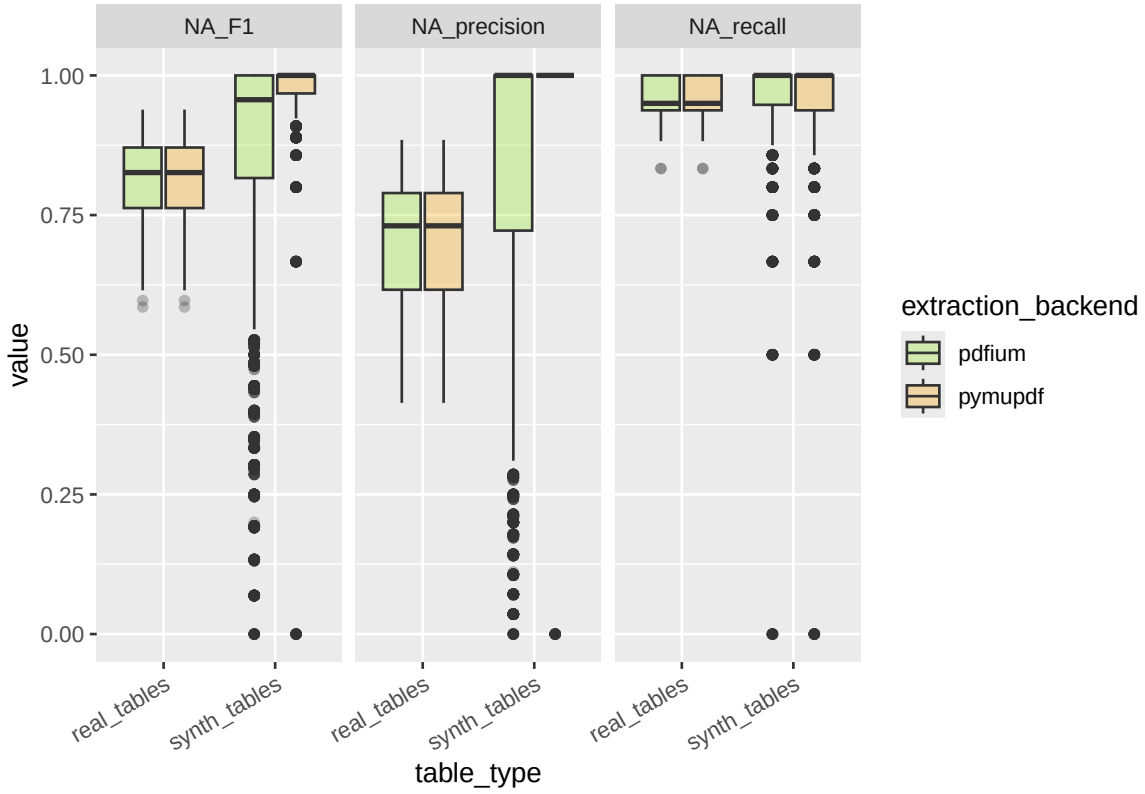


Figure 5.5: Performance on classification for missing values with regular expressions

1. The visual separation of columns or rows has an effect on the text processing at all.
2. It seems to have a positive effect on F1 and numeric correctness rate if the Passiva table is on the same page, even though it has no influence on the single predictions.

But one has to keep in mind that the number of data points on the aggregated values for the test set of the real dataset is only 18. So these findings are not strongly supporting any interpretation at all. Furthermore, the found effects are not very large - most below 5 %. Only the hypothesis for a positive influence of a missing of a value for the binomial prediction gets solid support with a mean absolute SHAP value of over 20‘ %. To get reliable results more tables have to be included which would require additional manual encoding.

```
htmltools::includeHTML(
  textConnection(
    xml2::read_html("../benchmark_results/table_extraction/hypotheses_and_results_real_table_extraction
      xml2::xml_find_first("//table") %>%
      as.character()
  )
)
```

5.2.1.0.1.2 Synthetic dataset Interpreting the visual results for the SHAP analysis on the synthetic dataset brought some interesting insides into the question under which condition the two PDF extraction libraries perform differently. These results can be treated as reliable since the model has been trained with 50_000 rows and the SHAP values have been calculated on 2_000 rows each.

Very interestingly the number of columns is having an opposite effect for the two libraries (see Figure 5.6A). Besides that often only pdfium struggled with some of the table characteristics while pymupdf is not influenced by them (for an example with header_span see Figure 5.6B).

It might be worth noting that the row for *Anteile an verbundenen Unternehmen* was rated to have a clear negative effect on the chance to extract the correct value.

Since there has no synthetic data created where also the Passiva table is present the result found with the real dataset can’t be investigated further. Also the question if visual separation is having an effect was not studied, even though, creating such additional synthetic data would be very easy with the current generation process and could be done in future work. It would be interesting if the visual separation is cause for the maleous text extractions of pdfium as well.

```
htmltools::includeHTML(
  textConnection(
    xml2::read_html("../benchmark_results/table_extraction/hypotheses_and_results_synth_table_extractio
      xml2::xml_find_first("//table") %>%
      as.character()
  )
)
```

5.2.2 Extraction with LLMs

- confidence usable to head for user checks?
- not handled new entries
- five examples bring not much more, but a little
- random forest / SHAP

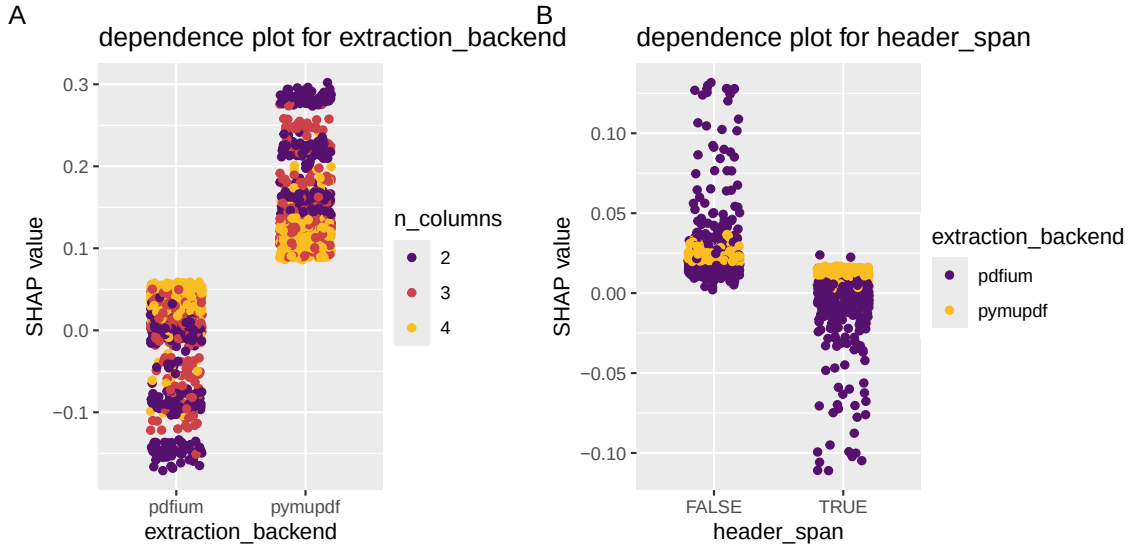


Figure 5.6: Showing the influence of the extraxtion library on the numeric text extraction task with synthetic data

5.2.2.1 Real tables only

For the table extraction task 30 open source models have been benchmarked[9]. The results are presented in Figure A.4, A.5 and A.6).

Most models need a context learning approach to beat the performance of the regular expression approach at total and numeric correctness rate and F1 score. Only 4 models perform better without any guidance[10] (see Table 5.5). 8 models achieved an performance better as the regex baseline using the approach to learn with a fixed example from the synthetic dataset.

In contrast: most of the models achieved a better performance than the regex baseline when they were provided with one or more examples from real *Aktiva* tables. Just 5 don't achieve a better value even with three or five realistic examples (see Table 5.6). Here we find the smallest models with less than 2B parameters which don't achieve a consistence performance no matter how many examples they get. But we also find models that start to perform bad if they get a too long context with too many examples like the very recent and large model Llama 4 Maverick.

With one and three examples the performance within one model family is positive correlated with the number of parameters the models have. Once the 4B parameters are passed the improvements get less and less getting closer to a perfect performance but never reaching it on all documents. Table 5.7 shows the mean performance for the best model-method approach for each model family. Most of the top performing model-method combinations rely on the maximum number of examples provided. Only the Llama-3 and Falcon3 model perform best with three examples[11].

Based on a small sample of 8 documents by the *Amt für Statistik Berlin-Brandenburg* it seems that there is support for the hypothesis, that providing Aktiva tables from the same company in in-context learning, is improving the results. This is especially noticeable for models with very few parameters and when providing only a single example. This seems intuitive, since there the potential for possibilities is much bigger. Figure A.7 shows that on this limited sample

[9] The models *deepseek-ai_DeepSeek-R1-Distill-Qwen-32B* and *google_gemma-3n-E4B-it* have been tested as well but don't get presented as they never performed anywhere beyond random guessing.

[10] There is an external guidance through the provided xgrammar template but it is not communicated to the model in a promt.

[11] Phi4 also perfroms best with three examples. But this is the maximum it can process due to a limited context length.

Table 5.5: Comparing table extraction performance with real 'Aktiva' dataset for models that perform well without or with little context learning

model	mean_total_zero_shot	mean_totalstatic_example
Llama4Maverick17B128EInstructFP8	0.717	0.747
Qwen2.532BInstruct	0.766	0.838
Qwen3235BA22BInstruct2507	{0.805}	{0.855}
phi4	0.792	0.703
MistralSmall3.124BInstruct2503	NA	0.788
Qwen2.572BInstruct	NA	0.765
Qwen330BA3BInstruct2507	NA	0.768
gemma327bit	NA	0.735

Table 5.6: Comparing table extraction performance with real 'Aktiva' dataset for models that worse than the regex baselin with 3 or 5 examples for incontext learning

model	method	mean_total
Llama4Maverick17B128EInstructFP8	top_5_rag_examples	0.004
Qwen2.50.5BInstruct	top_3_rag_examples_out_of_sample	0.094
Qwen2.51.5BInstruct	top_3_rag_examples_out_of_sample	0.611
Qwen30.6B	3_random_examples	0.443
gemma34bit	5_random_examples	{0.635}

Table 5.7: Comparing best mean table extraction performance with real 'Aktiva' dataset for each model family

model_family	model	method_family	n_examples	mean_total
Qwen 3	Qwen3235BA22BInstruct2507	top_n_rag_examples	5	0.961
Llama-4	Llama4Scout17B16EInstruct	top_n_rag_examples	5	0.931
mistralai	MistralLargeInstruct2411	top_n_rag_examples	5	0.929
Llama-3	Llama3.170BInstruct	n_random_examples	3	0.911
Qwen 2.5	Qwen2.514BInstruct	n_random_examples	5	0.908
microsoft	phi4	n_random_examples	3	0.893
tiiuae	Falcon310BInstruct	top_n_rag_examples	3	0.862
google	gemma327bit	n_random_examples	5	0.821

- the improvement is bigger for Qwen 3 than for Qwen 2.5
- Googles gemma 27b and GPT 4.1 mini could overcome an unnoticed issue with the extraction with just one example.
- the effect of being overwhelmed by a too rich context with LLamas Maverick model could get reduced a bit.

To examine the question, if the reported confidence score of the responses can be used, to flag the predicted values as potentially wrong. Again, Figure 5.7 shows, that Qwen 3 reports very high confidence values no matter if the results are correct or not. With the Mistral model we find a wider range of confidences given and for wrong results lower confidence is reported.

Figure 5.8 shows, that the chance to make an mistake by believing the prediction is rising with lower confidence. The chance to make a mistake is higher for predictions of numeric values than for believing a value is not present in the table. The chance to make such a mistake is higher using the confidence reported by Qwen 3.

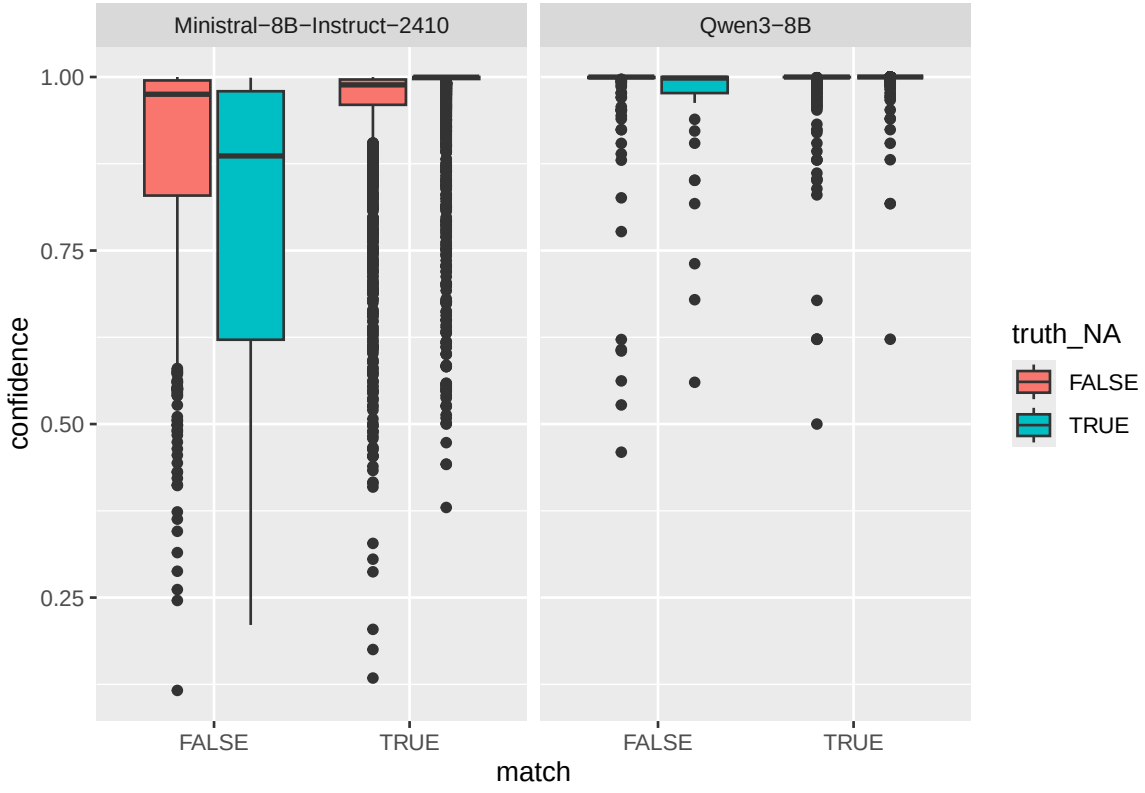


Figure 5.7: Comparing the reported confidence scores for the table extraction task on real dataset for the Mistral and Qwen 3 with 8B parameters.

5.2.2.1.1 Hypotheses The formulated hypotheses have been evaluated visually using the dependence and beeswarm plots from the shapviz library based on the SHAP values calculated with a random forest.

5.2.2.1.1.1 Real dataset Even though the samples size of Aktiva tables did not increase, the available training, test and SHAP sample size is much larger, because the experiment has been repeated with different models and methods. Thus, the interpretations based on the visual evaluation (see Figure A.2)) are more reliable for model and method specific predictors. Since there is one Aktiva example for every company files

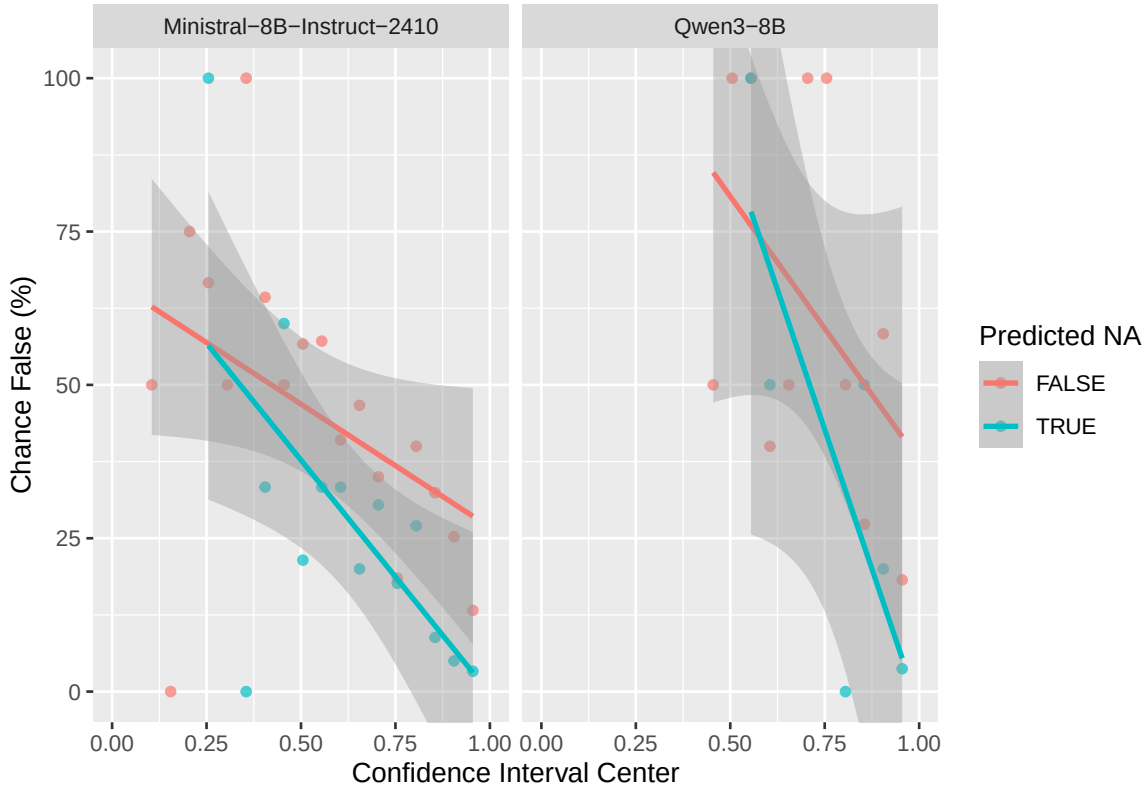


Figure 5.8: Estimating the relative frequency to find a wrong extraction result over different confidence intervals

were found for they might even be generalizable for this population. But one has to keep in mind that there have been more Aktiva tables for *Amt Stat BBB* which might nudge the results a bit.

The results assign much more influence on model and method specific attributes than on the table specific attributes. The importance of the table attributes are as low as found with the regular expresion approach. Only for the binomial prediction we find the predictor *missing* to get assigned more importance than to all model and method specific attributes. Same is true for the *label* that is having the highest influence on the reported confidence. Nevertheless, in the case of the binomial prediction there is half of the predictors *missing* and *label* importance shifted to model and method specific predictors.

Again, multiple hypotheses don't get supported by the visual results. The surprising results are:

1. In general more examples are helpful except for Llamas Maverick model that performs poorly with five examples. But this effect is only noticeable with the aggregated metrics nor for the case wise binomial evaluation.
2. The number of columns has a negative effect on the performance but no effect on the reported confidence.
3. There was no negative effect found if the *Passiva* table is on the same page as the Aktiva table.
4. Larger models start to report less confidence again. This is not unexpected for the Mistral model but was surprising for the largest Qwen 3 model. (Discussion: New Generation? Aktive paramaters count? Irrelevant because not well distinguishing?)
5. It not only influences the the performance to extract the correct numeric value from a row where there are additional sums present but also the F1 score.

Two interesting details found while inspecting the dependence plots for the metric *percentage_numeric_correct* are (see Figure 5.9A) that the bad performance of LLamas Maverick with five examples is easily spottable and that the negative effect of *T_in_year* might be caused by an interaction with *vis_separated_rows* completely (see Figure 5.9B). To investigate the second finding one would need tables where the uni is present in the year column and having no visual separation of the rows at the same time. Synthetic data potentiall could help to answer such questions.

```
htmltools::includeHTML(
  textConnection(
    xml2::read_html("../benchmark_results/table_extraction/hypotheses_and_results_real_table_extraction_l
      xml2::xml_find_first("//table") %>%
      as.character()
  )
)
```

5.2.2.1.2 GPT Even though a lot of documents to processed at RHvB (Rechnungshof von Berlin) will not be public and thus must not be processed on public cloud infrastructure, the performance of models like OpenAI's GPT or Google's Gemini are interesting benchmark references within this thesis and for comparing these findings with other papers results. Therefore for this thesis the public available versions of annual reports have been used instead of the ones used internally or for public administration purposes. Those public available reports often are visually more appealing and more heterogeneous in their structure.

As a reference to compare the performance of OpenAI's models with the results of four Qwen 3 models are shown as well. Surprisingly gpt-5-mini performed best among all models of OpenAI and is performing as good as the top Qwen 3 model.

Using gpt-oss-20b, gpt-5-nano and gpt-5-chat for the table extraction task was not working. With gpt-5-nano the answers were not respecting the provided grammar. Running gpt-5-chat resultet in the error informing that a *json_schema* can't be used with this model. With gpt-5-mini the very approach worked flawless. Running gpt-oss-20b with the vllm offline inference framework was possible but the xgrammar was not respected

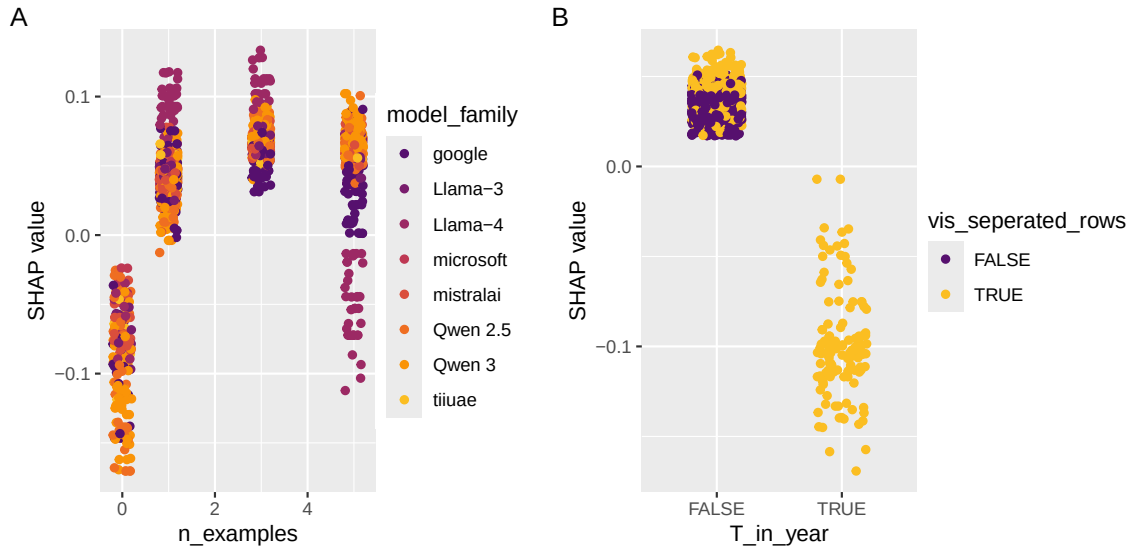


Figure 5.9: Showing the influence of many examples on Llama 4 Maverick (A) and interaction between *T in year* and *vis separated rows* (B)

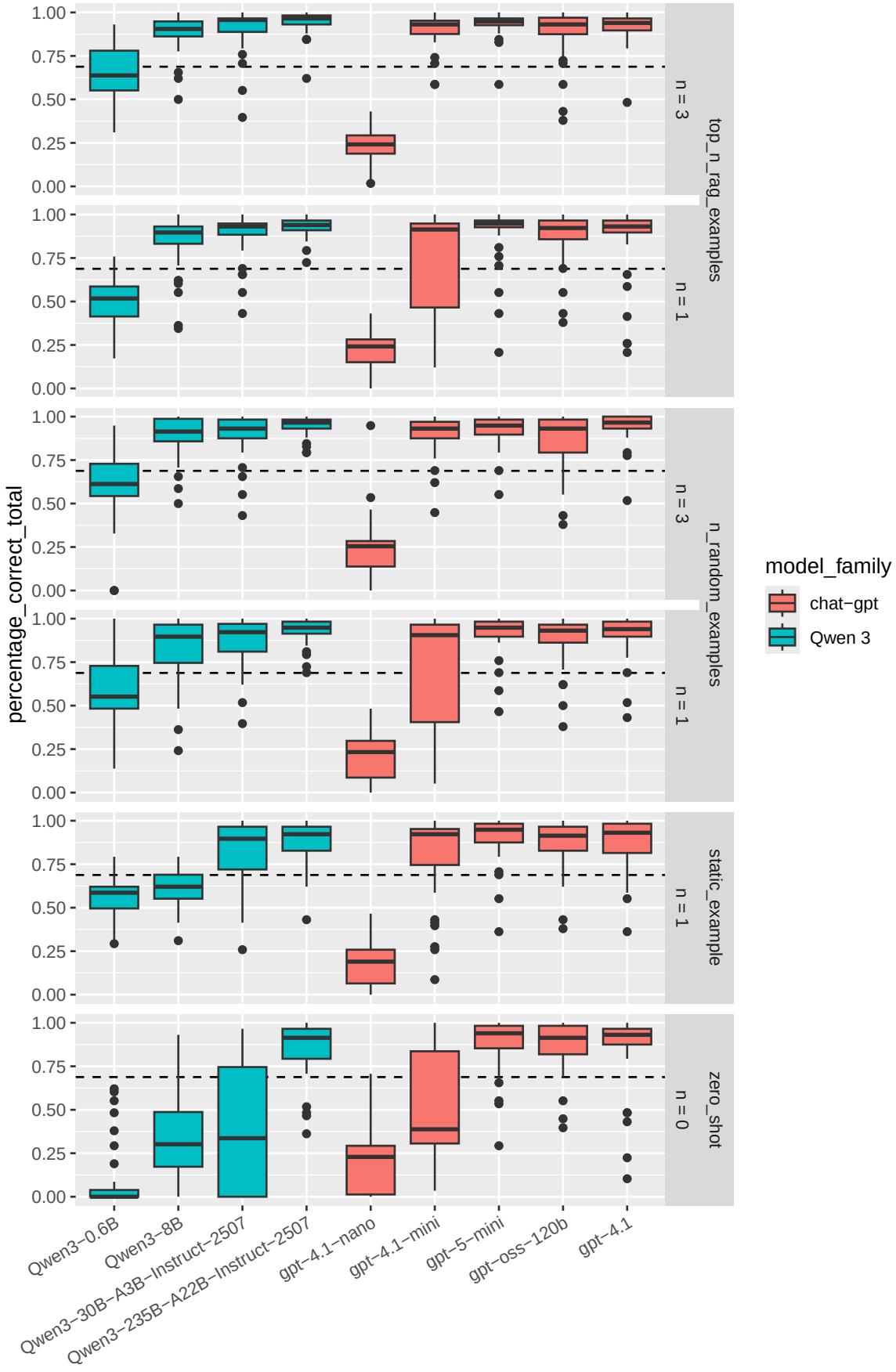
by the new harmony output format. With a gpt-oss-120b instance hosted on Azure the guided decoding worked.

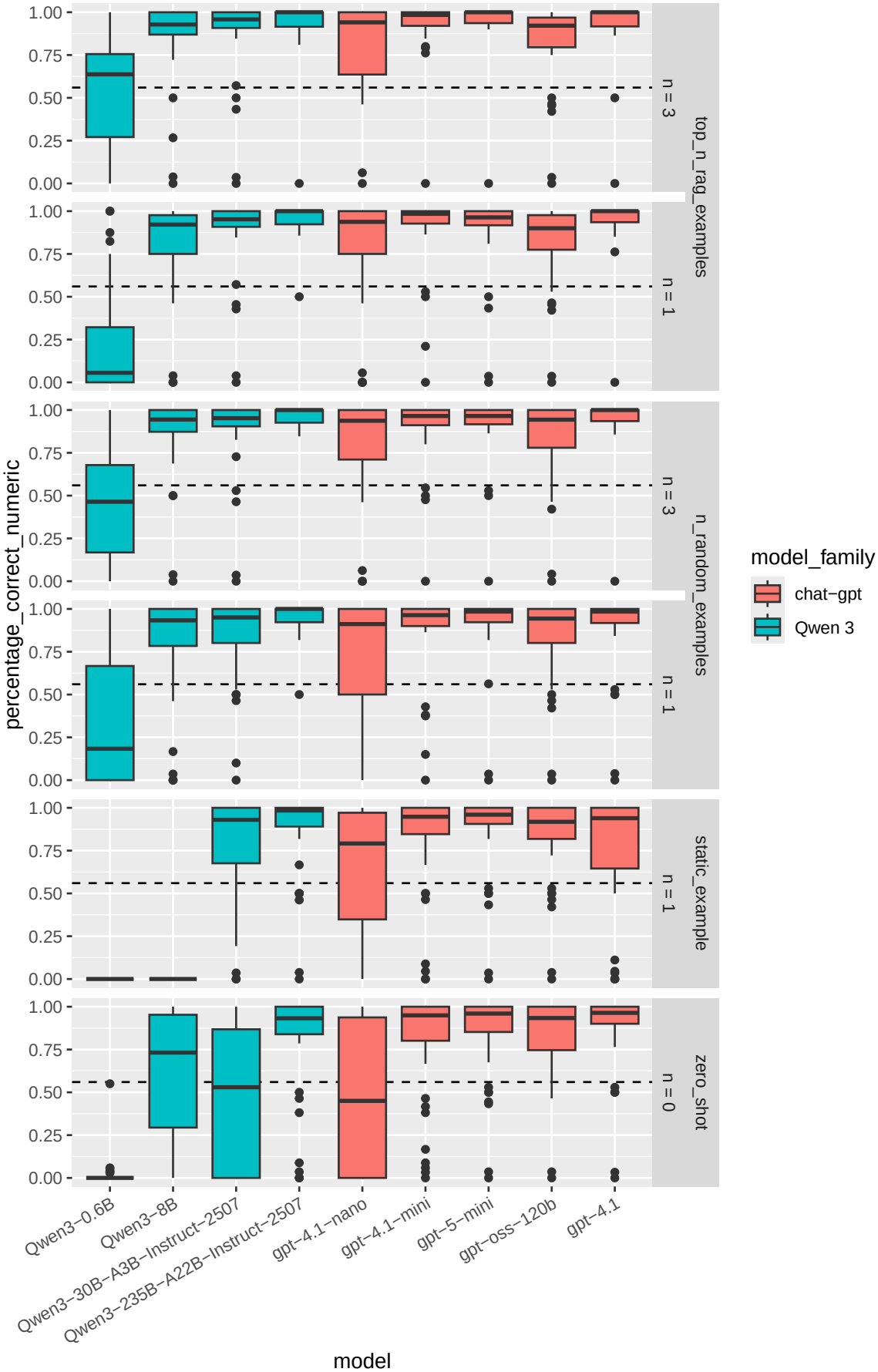
GPT 4.1 nano not sufficiently powerful (can't extract numerics, but NA prediction okay), GPT 4.1 mini good to moderate in most cases, GPT 4.1 performs very good. How does gpt-4.1-nano s total score can be so low? Predicting not all rows and then just checking present rows?

GPT 4.1 costs five times of mini and 20 times of nano. But nano is useless for the task

GPT oss 120 b better as 4.1 mini, GPT 5 mini better (best) as 4.1

Costs for gpt-5-mini and gpt-oss-120b not shown in Azure yet. :(


percentage_correct_total
1.00
0.75
0.50
0.25
0.00
top_n_rag_examples
n = 3
n = 1
n_random_examples
n = 3
n = 1
static_example
n = 1
zero_shot
n = 0
model_family
chat−gpt
Qwen 3
Qwen3−0.6B
Qwen3−8B
Qwen3−30B−A3B−Instruct−2507
Qwen3−235B−A22B−Instruct−2507
gpt−4.1−nano
gpt−4.1−mini
gpt−5−mini
gpt−oss−120b
gpt−4.1
model


percentage_correct_numeric
1.00
0.75
0.50
0.25
0.00
top_n_rag_examples
n = 3
top_n_rag_examples
n = 1
n_random_examples
n = 3
n_random_examples
n = 1
static_example
n = 1
zero_shot
n = 0
Qwen3−0.6B
Qwen3−8B
Qwen3−30B−A3B−Instruct−2507
Qwen3−235B−A22B−Instruct−2507
gpt−4.1−nano
gpt−4.1−mini
gpt−5−mini
gpt−oss−120b
gpt−4.1
model
model_family
chat−gpt
Qwen 3

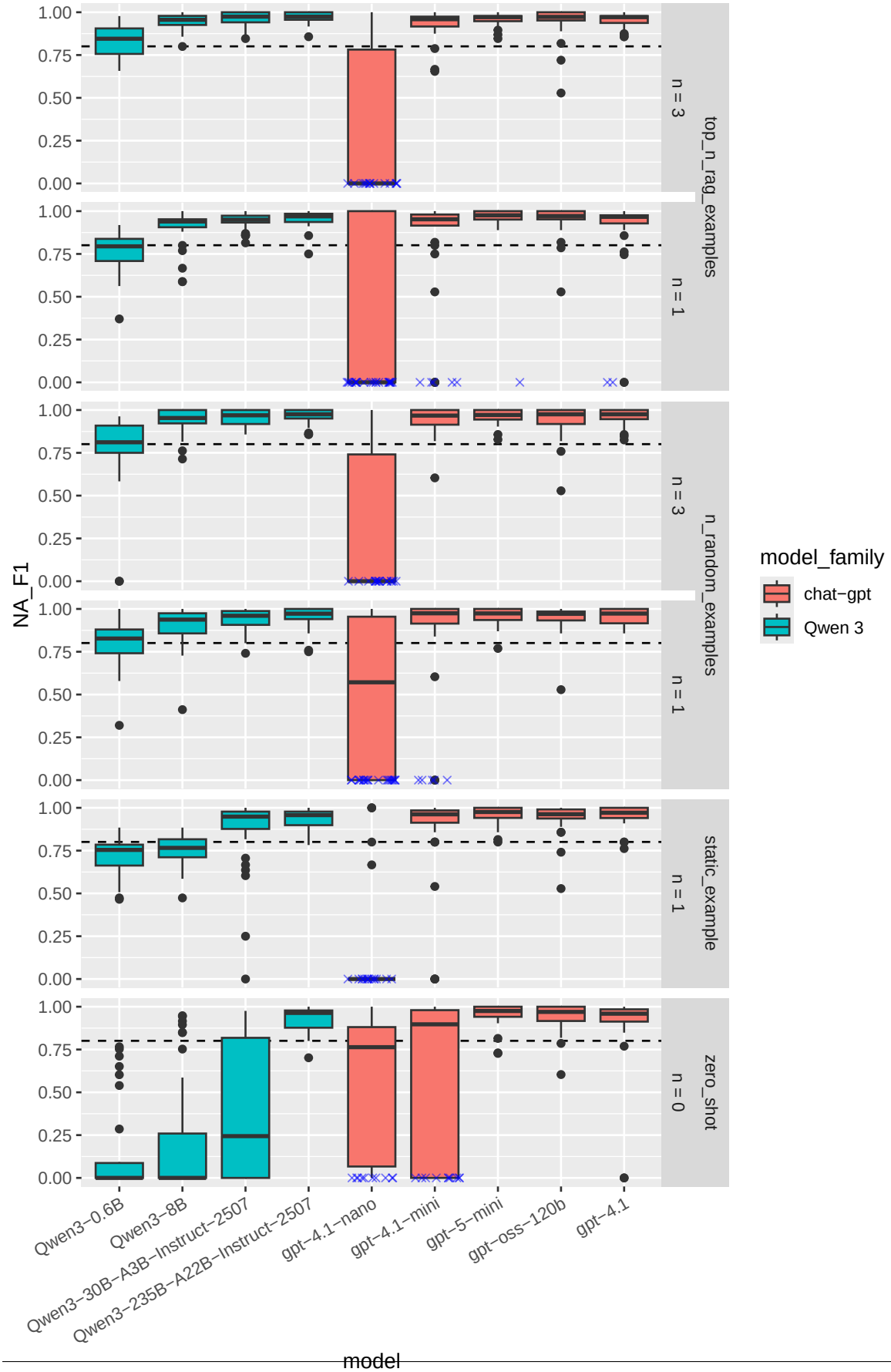


Figure 5.10: The blue crosses indicate runs where a model has predicted only numeric values even though there have been missing values.

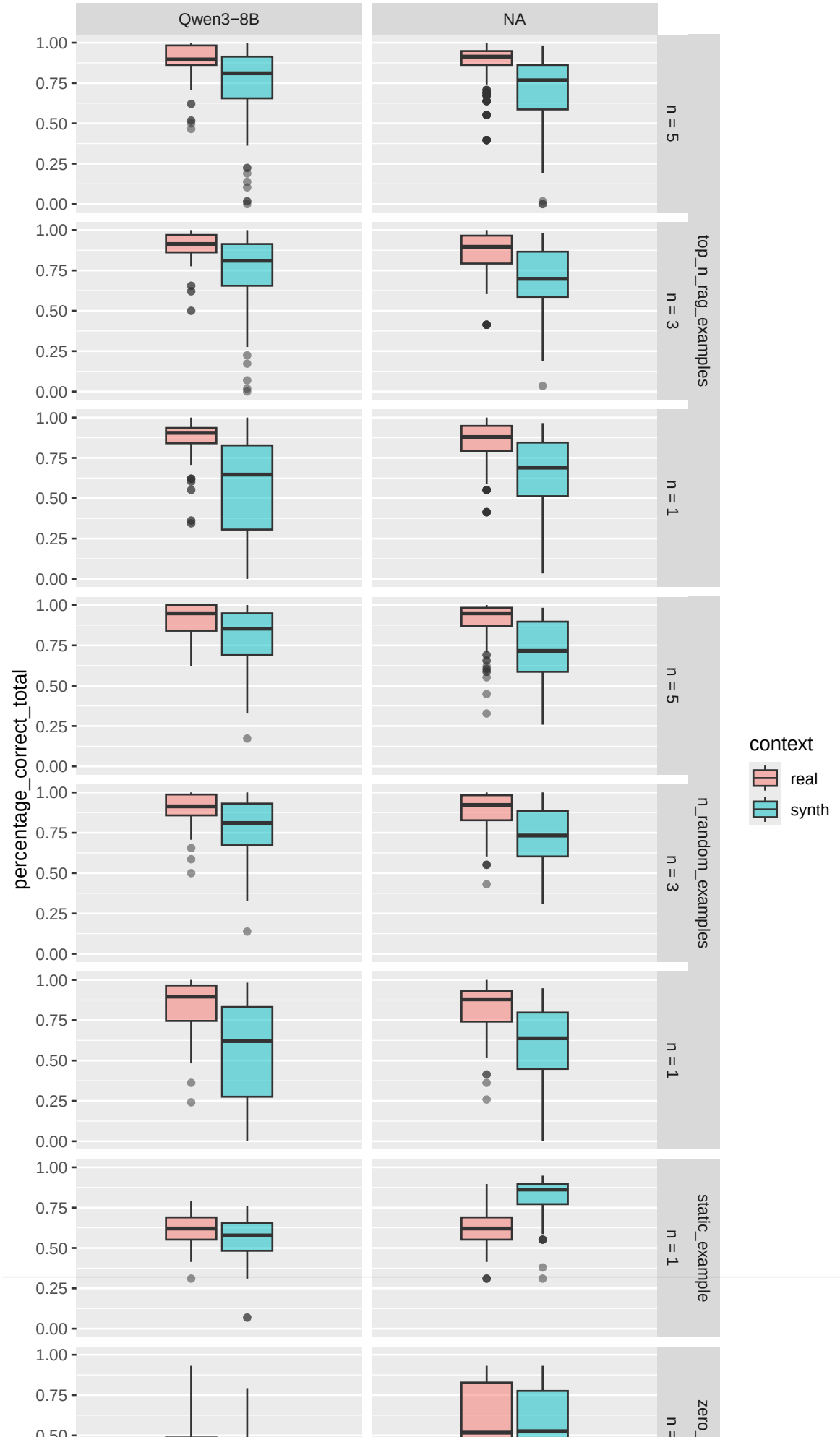
Meter	Cost	Currency	Cost_all_tasks
gpt 4.1 Inp glbl Tokens	3.53	EUR	7.02
gpt 4.1 Outp glbl Tokens	2.71	EUR	3.44
gpt 4.1 mini Inp glbl Tokens	1.23	EUR	1.23
gpt 4.1 mini Outp glbl Tokens	0.71	EUR	0.71
gpt 4.1 nano Inp glbl Tokens	0.31	EUR	0.31
gpt 4.1 nano Outp glbl Tokens	0.15	EUR	0.15

5.2.2.2 Synthetic tables only

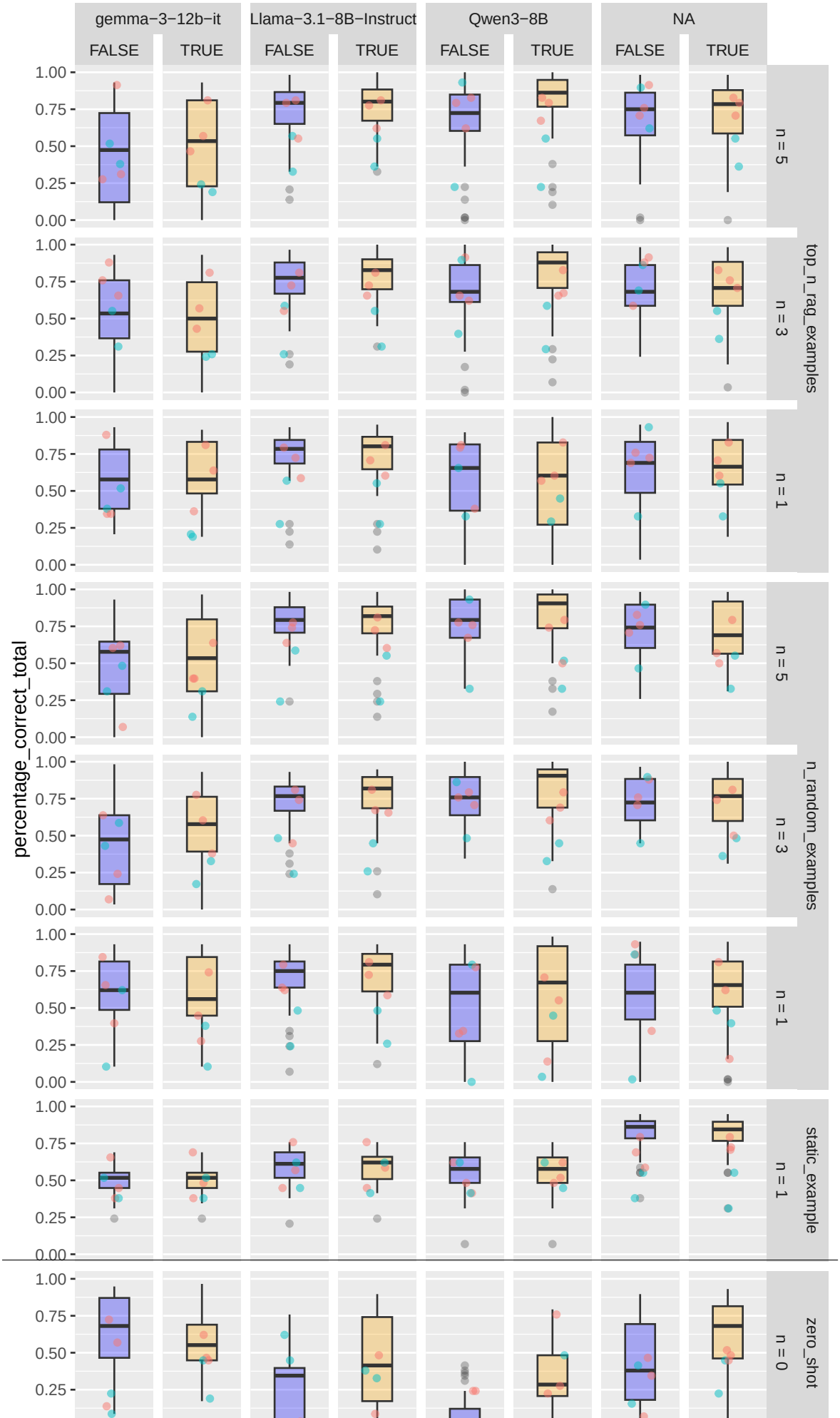
span argument was not implemented correct in html tables and md :/

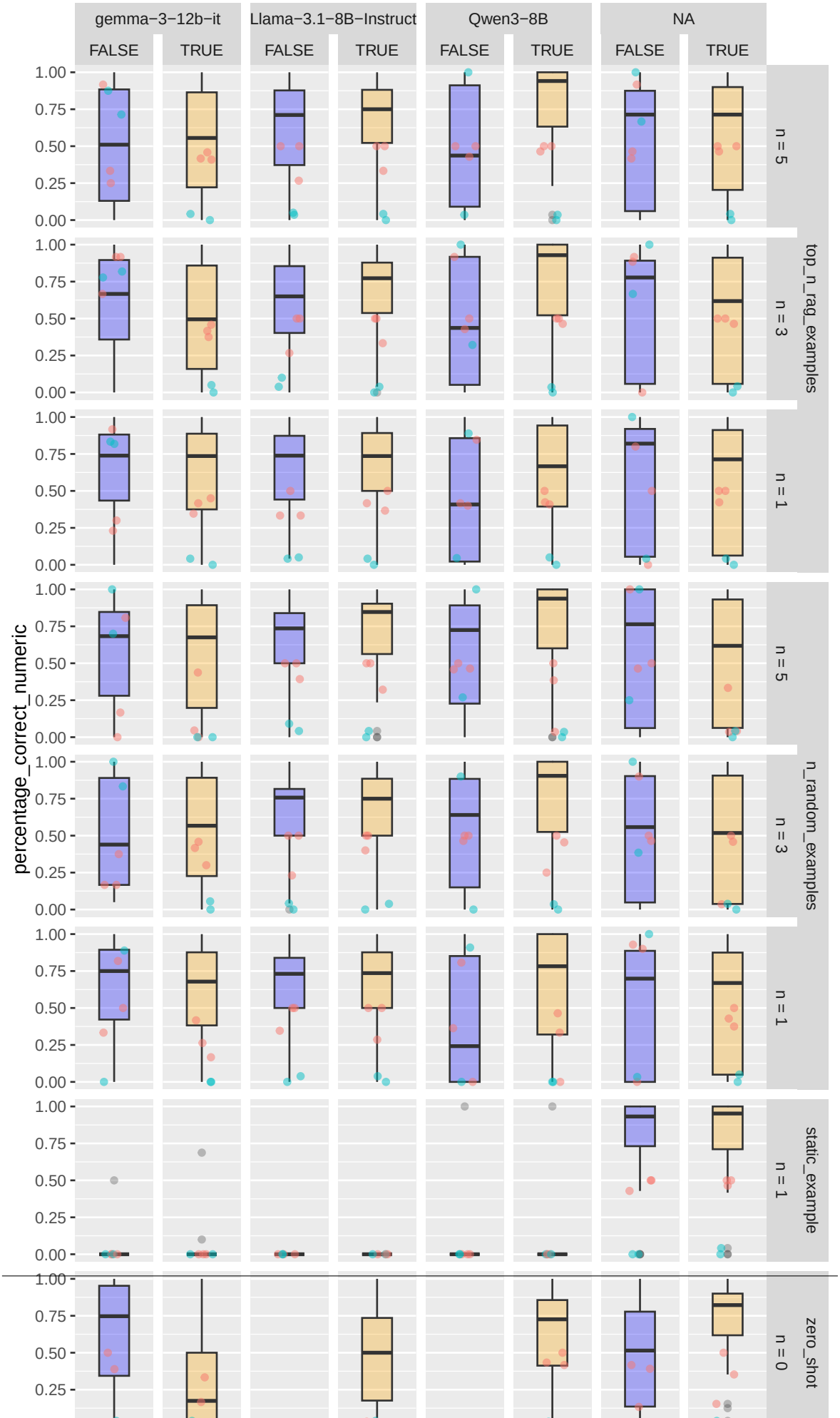
5.2.2.3 Extract from real tables with synthetic content

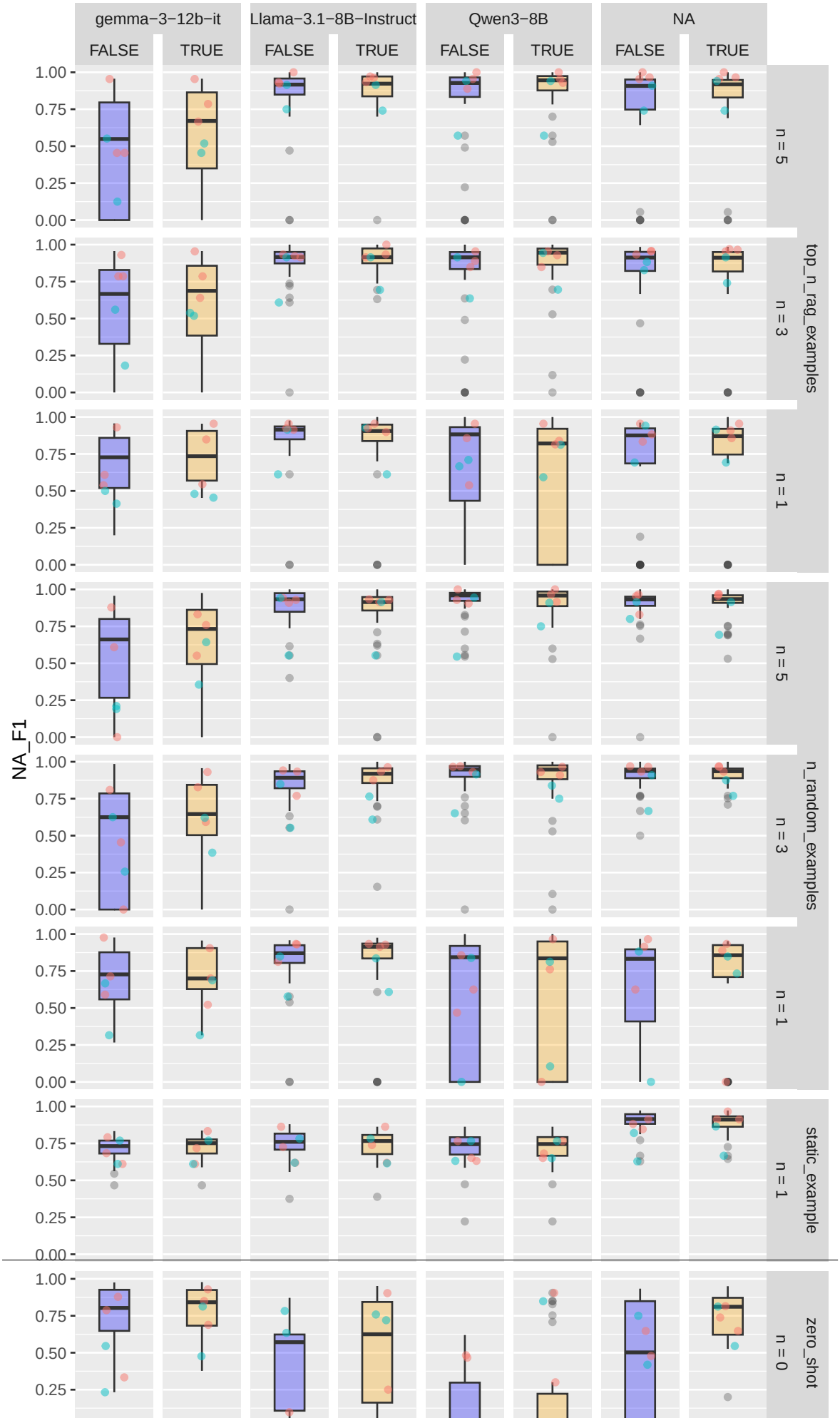
Real context better for real tables. But not useless.

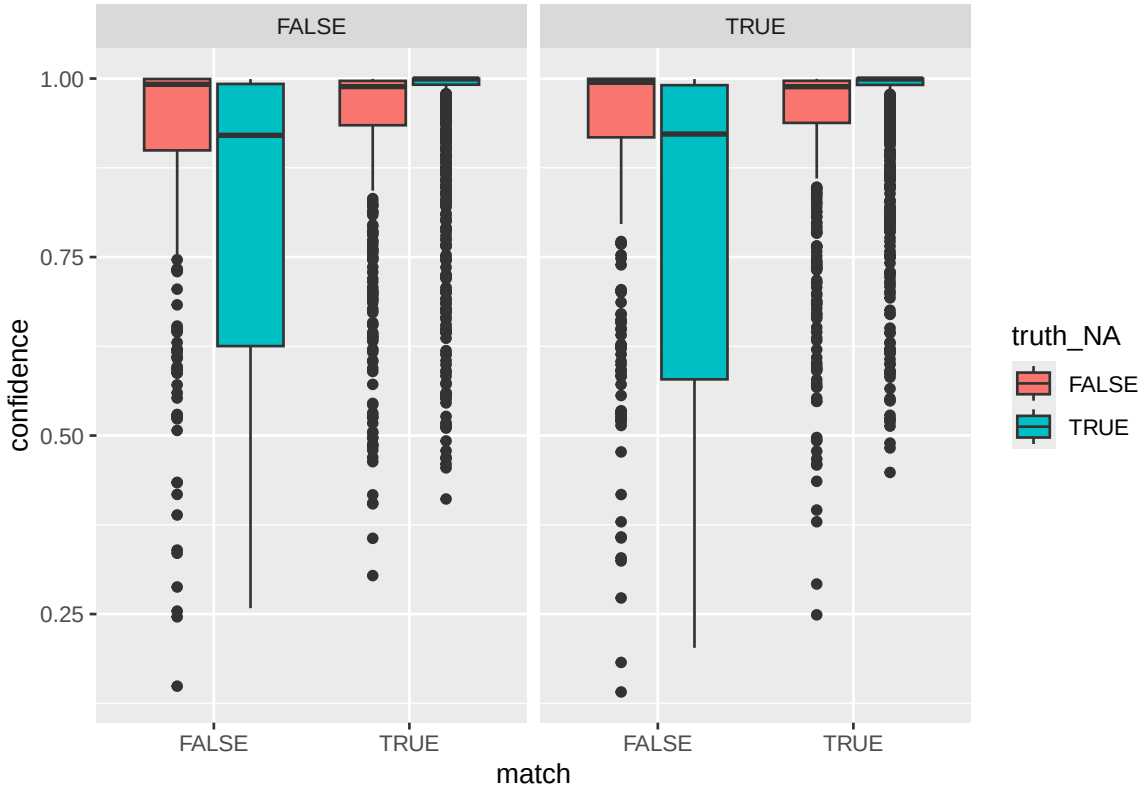

Qwen3−8B
NA
top_n_rag_examples
n_random_examples
static_example
zero_
n = 5
n = 3
n = 1
percentage_correct_total
1.00
0.75
0.50
0.25
0.00
context
real
synth

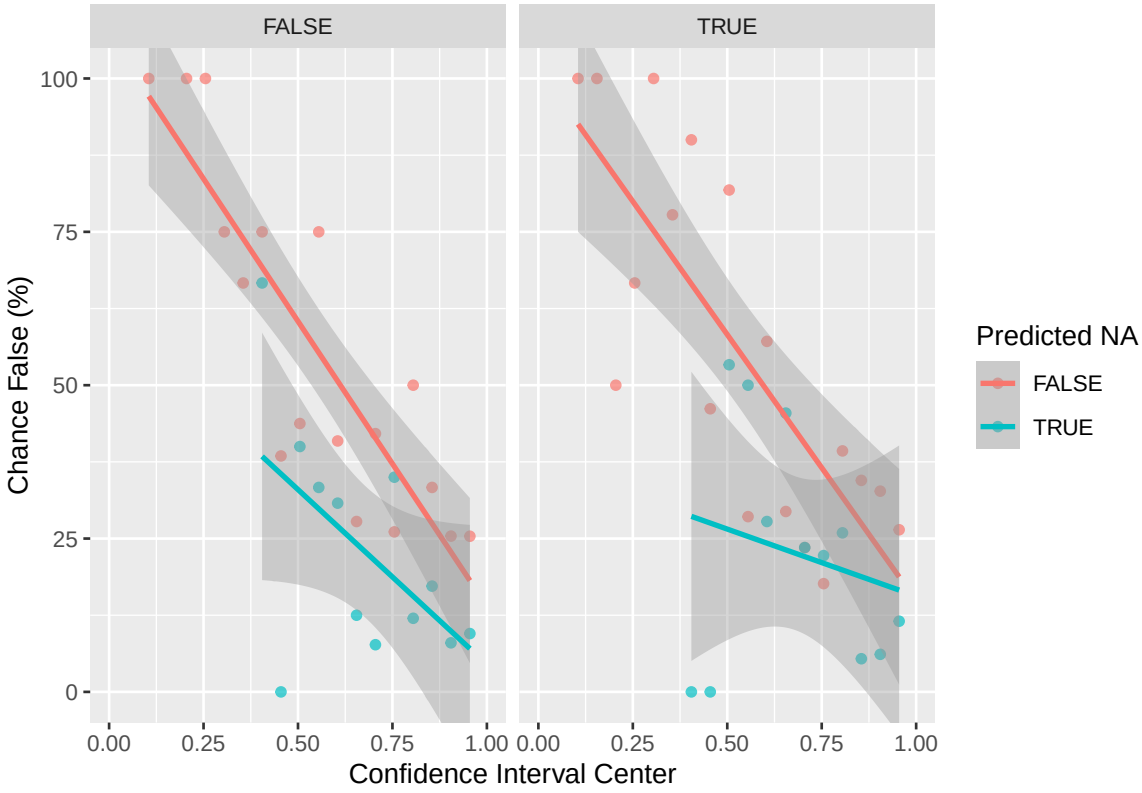
no examples with units only for one column. Can learn from synth context new skills


gemma−3−12b−it
Llama−3.1−8B−Instruct
Qwen3−8B
NA
FALSE
TRUE
percentage_correct_total
1.00
0.75
0.50
0.25
0.00
top_n_rag_examples
n_random_examples
static_example
zero_shot
n = 5
n = 3
n = 1
n = 0


gemma−3−12b−it
Llama−3.1−8B−Instruct
Qwen3−8B
NA
FALSE
TRUE
percentage_correct_numeric
1.00
0.75
0.50
0.25
0.00
top_n_rag_examples
n_random_examples
static_example
zero_shot
n = 5
n = 3
n = 1
n = 0


gemma−3−12b−it
Llama−3.1−8B−Instruct
Qwen3−8B
NA
FALSE
TRUE
NA_F1
1.00
0.75
0.50
0.25
0.00
top_n_rag_examples
n_random_examples
static_example
zero_shot
n = 5
n = 3
n = 1
n = 0


FALSE
TRUE
1.00
0.75
0.50
0.25
confidence
FALSE
TRUE
FALSE
TRUE
match
truth_NA
FALSE
TRUE


FALSE
TRUE
100
75
50
25
0
Chance False (%)
0.00
0.25
0.50
0.75
1.00
Confidence Interval Center
Predicted NA
FALSE
TRUE

5.2.3 Comparison

Chapter 6

Discussion

- ensemble prediction
- check for halluzination vs wrong placed / repeated numbers

6.1 Table extraction

building a document extraction database document by document can improve performance taking advantage of same-company rag in-context learning

6.1.1 Regex baseline

- synthetic tables have been generated with cell lines because this should have improved the performance of a table extraction approach (not conducted)- maybe this is confusing pdfium? Or the zoom level?

6.2 Not covered

- OCR
- fine-tuning
- using something smaller (e.g. LSTMs) instead LLMs
- building application, UX design (ref Ambacher 2024)
- table extraction (either VLLMs or classic approaches <– tyred tabula but was not successful (because of missing visual traits)?)

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.*

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: An Overview.* arXiv. https://doi.org/10.48550/arXiv.2008.05756

Li, H., Gao, H. (Harry)., Wu, C., & Vasarhelyi, M. A. (2023). *Extracting Financial Data from Unstructured Sources: Leveraging Large Language Models* [{SSRN} {Scholarly} {Paper}]. Social Science Research Network. https://doi.org/10.2139/ssrn.4567607

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

List of Figures

List of Tables

Glossary

LLM large language model
RHvB Rechnungshof von Berlin
Rmd Rmarkdown document
SHAP SHapley Additive exPlanations
YAML YAML Ain't Markup Language
glm generalized linear model
regex regular expression

Chapter A

Appendix

A.1 Local machine

One can find the specifications of the local machine used to run the less computationally demanding tasks below. It is a lightweight laptop device. Its performance cores support hyperthreading and have a clock range between 2.1 and 4.7 GHz. However, due to the flat design, there is little active cooling. Thus, thermal throttling starts rather quickly. It is therefore a reasonable assumption that most locally benchmarked tasks are running at 2.1 GHz. Despite this handicap, it has a sufficiently large RAM of 32 GB and 3 GB of NVMe disk space.

System Details Report

Report details

- **Date generated:** 2025-07-19 13:56:16

Hardware Information:

- **Hardware Model:** LG Electronics 17ZB90Q-G.AD79G
- **Memory:** 32.0 GiB
- **Processor:** 12th Gen Intel® Core™ i7-1260P × 16
- **Graphics:** Intel® Graphics (ADL GT2)
- **Disk Capacity:** 3.0 TB

Software Information:

- **Firmware Version:** A2ZG0150 X64
- **OS Name:** Ubuntu 24.04.2 LTS
- **OS Build:** (null)
- **OS Type:** 64-bit
- **GNOME Version:** 46
- **Windowing System:** Wayland
- **Kernel Version:** Linux 6.11.0-29-generic

Table A.1: Comparing extraction time (in seconds) for different Python package

package	runtime in s
pdfium	{14}
pymupdf	22
pypdf	218
pdfplumber	675
pdfminer	752
doclingparse	1621

A.2 Benchmarks

A.2.1 Text extraction

A basic requirement for all succeeding tasks is, that the text gets extracted from the PDF files. As written in doclings technical report (Auer et al., 2024) the available open source libraries differ in their speed and restrictiveness of licensing. Since there are no benchmark results this report multiple libraries have been tested here.

The benchmark ran on the local machine described in section A.1. There have been 5256 pages to extract the text from.

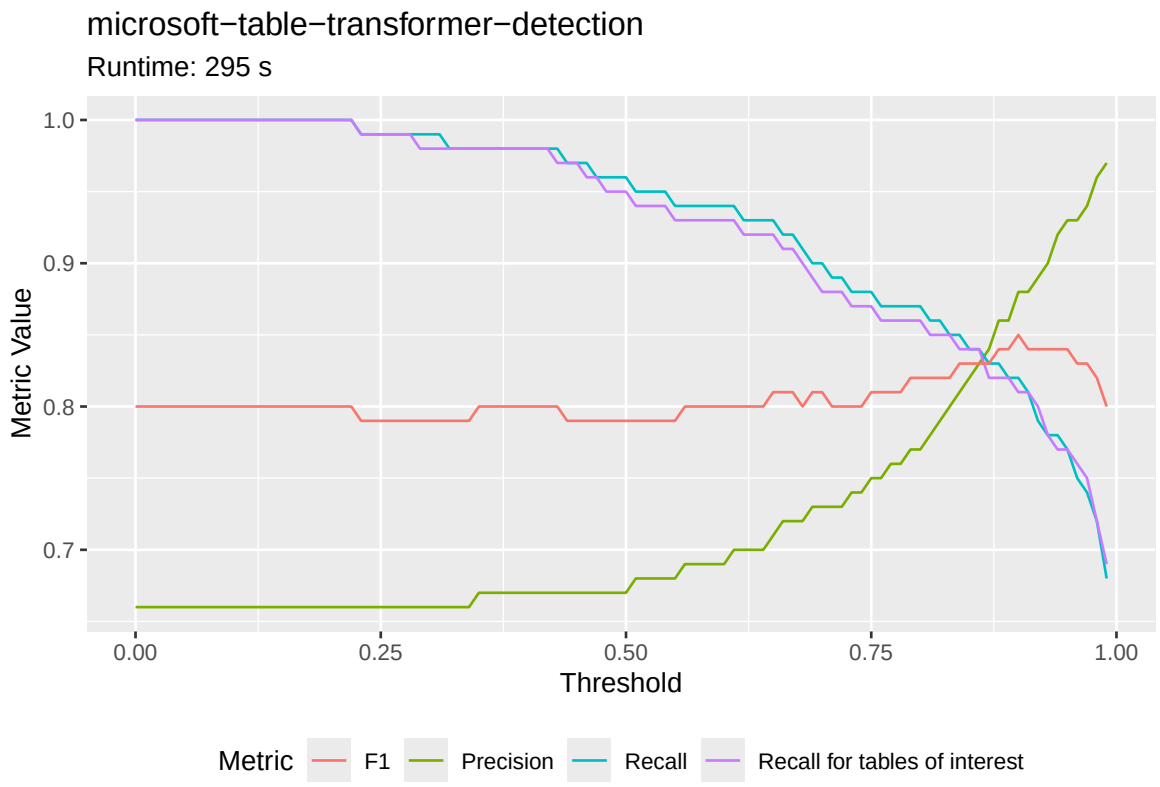
The result of docling-parse is not formated as markdown yet but also just plain text.

For implementation in a system where the text has to get extracted live or frequently the speed of the library might be paramount. But in special cases it can be important to invest more computational power into text extraction if this assures extraction according a more complicated document layout. E.g. some of the tables have been parsed by pdfium in such a manner that first all row descriptors have been extracted (first row) and thereafter all numeric columns (rowwise) ADD REFERENCE / EXAMPLE.

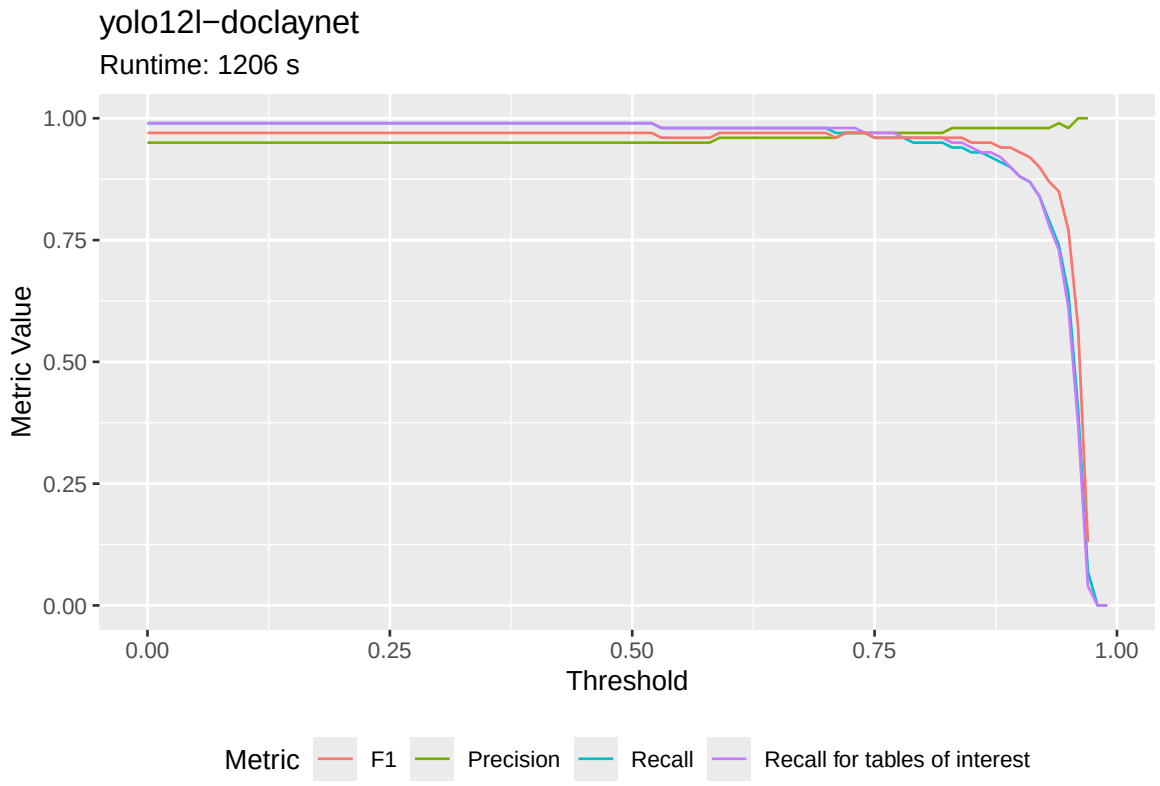
A.2.2 Table detection

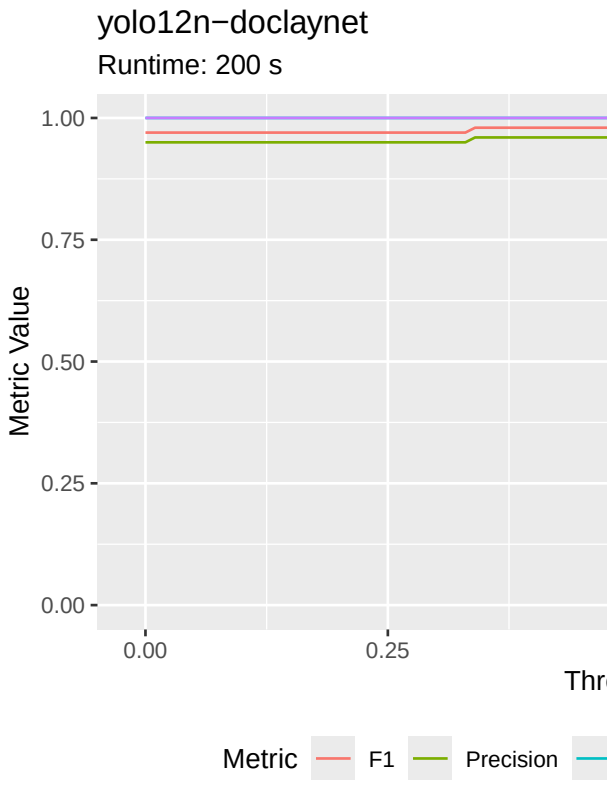
- yolo benchmark and table transformer
- skip classification with llm

not so important anymore

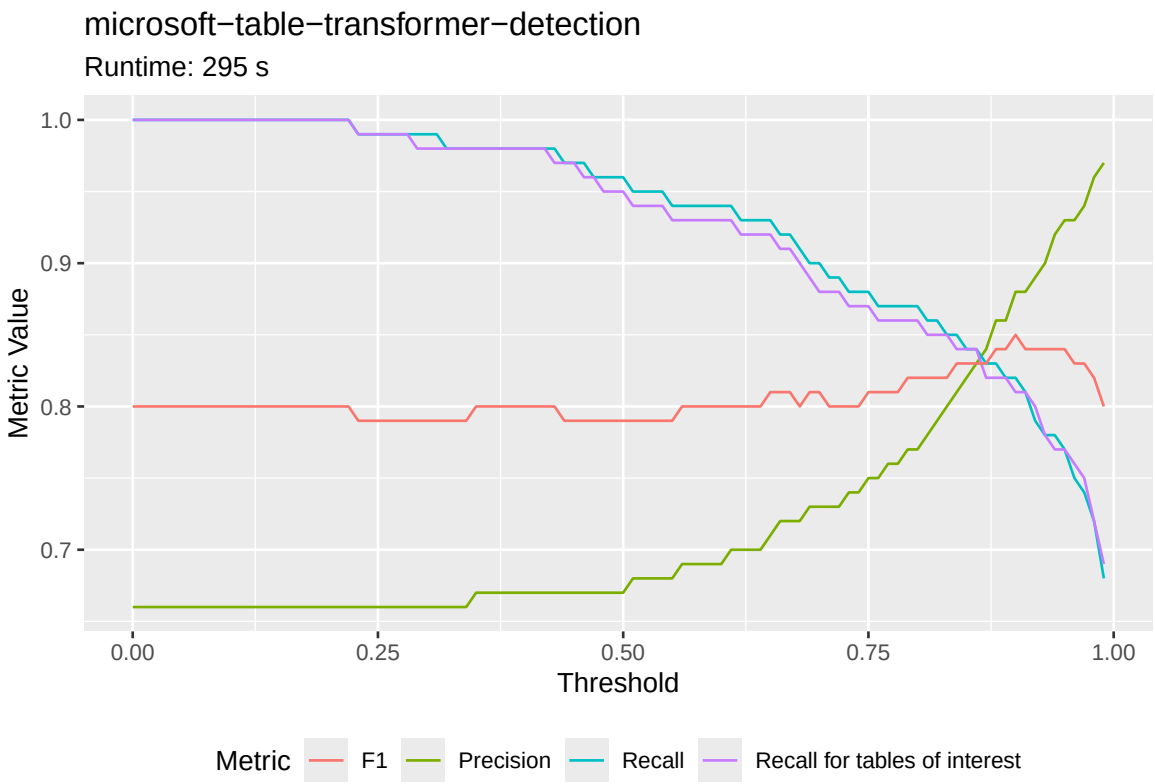


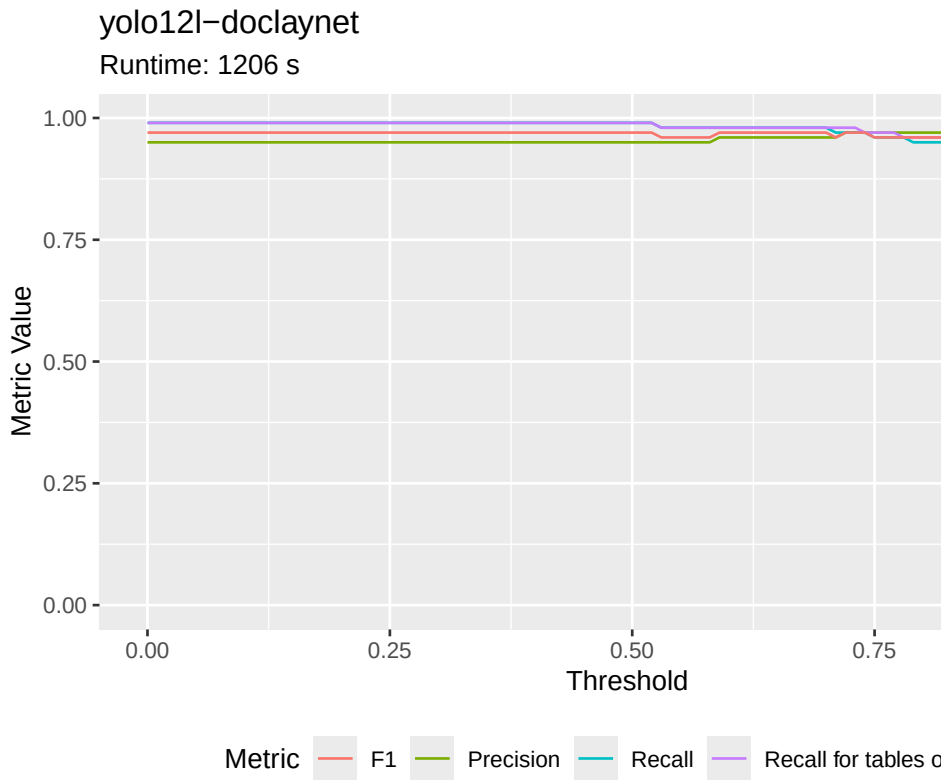
You see the plot for: microsoft-table-transformer-detection. (Click to stop automatic rotation.)



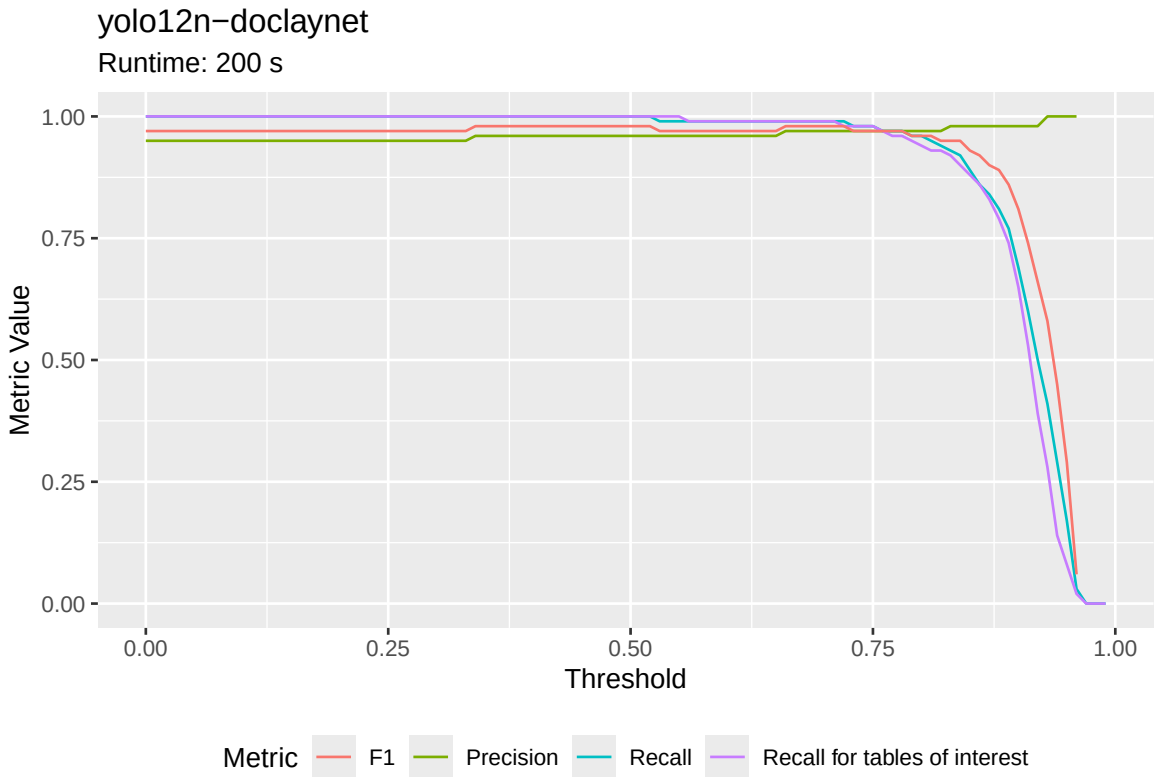


You see the plot for: yolo12l-doclaynet. (Click to stop automatic rotation.)
You see the plot for: yolo12n-doclaynet. (Click to stop automatic rotation.)





microsoft-table-transformer-detection



yolo12l-doclaynet
yolo12n-doclaynet

Table A.2: Comparing time (in seconds) for processing ten asset tables using different libraries and approaches

Model parameters (in B)	Transformers	vLLM	vLLM batched
0.5	330	65	NA
3.0	628	130	20
7.0	940	217	30

A.2.3 Large language model process speed

In April 2025 there have been issues with running vllm within the Python framework. Thus the first experiments have been conducted using the transformers library. When the problems of building a working vllm based docker image for the experiments it was measured how long the same task takes with the transformers and the vllm library and how the batched processing competes versus a loop approach. The model family used was Qwen 2.5 Instruct. The task was to extract the assets table for ten real example pages.

Table A.2 shows that the experiments with vllm library run are around four to five times faster. Processing the messages in a batched mode again is six to seven times faster.

The change of the experimental setup from transformers loop-based to vllm batched mode made is possible run the benchmark on whole PDF documents giving a sound estimate of the false positive rate in the page identification task (see section 5.1.3). Previous experiments have only been using a subset of pages that have been selected with the baseline regex approach (see section 5.1.1).

A.3 Regular expressions

Here one can find the three regular expressions used for the benchmarks presented in section 5.1.1.

```
simple_regex_patterns = {
    "Aktiva": [
        r"aktiva",
        r"((20\d{2}).*(20\d{2}))"
    ],
    "Passiva": [
        r"passiva",
        r"((20\d{2}).*(20\d{2}))"
    ],
    "GuV": [
        r"gewinn",
        r"verlust",
        r"rechnung",
        r"((20\d{2}).*(20\d{2}))"
    ]
}
```

```
regex_patterns_5 = {
    "Aktiva": [
        r"a\s*k\s*t\s*i\s*v\s*a|a\s*k\s*t\s*i\s*v\s*s\s*e\s*i\s*t\s*e|anlageverm.{1,2}gen",
        r"((20\d{2}).*(20\d{2}))|((20\d{2}).*vorjahr)|vorjahr",
        r"Umlaufverm.{1,2}gen|Anlageverm.{1,2}gen|Rechnungsabgrenzungsposten|Forderungen",
        r"\s([a-zA-Z]|[0-9]{1,2}|[iI]+)[\.\)]\s"
    ],
    "Passiva": [
```

```
        r"p\s*a\s*s\s*s\s*i\s*v\s*a|p\s*a\s*s\s*s\s*i\s*v\s*s\s*e\s*i\s*t\s*e|eigenkapital",
        r"((20\d{2}).*(20\d{2}))|((20\d{2}).*vorjahr)|vorjahr",
        r"Eigenkapital|R.{1,2}ckstellungen|Verbindlichkeiten|Rechnungsabgrenzungsposten",
        r"\s([a-zA-Z]|[0-9]{1,2}|[iI]+)[\.\)]\s"
    ],
    "GuV": [
        r"gewinn|guv",
        r"verlust|guv",
        r"rechnung|guv",
        r"((20\d{2}).*(20\d{2}))|vorjahr"
        r"Umsatzerl.{1,2}se|Materialaufwand|Personalaufwand|Abschreibungen|Jahres.{1,2}berschuss|Jahres
        r"\s([a-zA-Z]|[0-9]{1,2}|[iI]+)[\.\)]\s"
    ]
}
```

```
regex_patterns_3 = {
    "Aktiva": [
        r"a\s*k\s*t\s*i\s*v\s*a|a\s*k\s*t\s*i\s*v\s*s\s*e\s*i\s*t\s*e|anlageverm.{1,2}gen",
        r"((20\d{2}).*(20\d{2}))|((20\d{2}).*vorjahr)|vorjahr"
    ],
    "Passiva": [
        r"p\s*a\s*s\s*s\s*i\s*v\s*a|p\s*a\s*s\s*s\s*i\s*v\s*s\s*e\s*i\s*t\s*e|eigenkapital",
        r"((20\d{2}).*(20\d{2}))|((20\d{2}).*vorjahr)|vorjahr"
    ],
    "GuV": [
        r"gewinn|guv",
        r"verlust|guv",
        r"rechnung|guv",
        r"((20\d{2}).*(20\d{2}))|vorjahr"
    ]
}
```

A.4 Figures

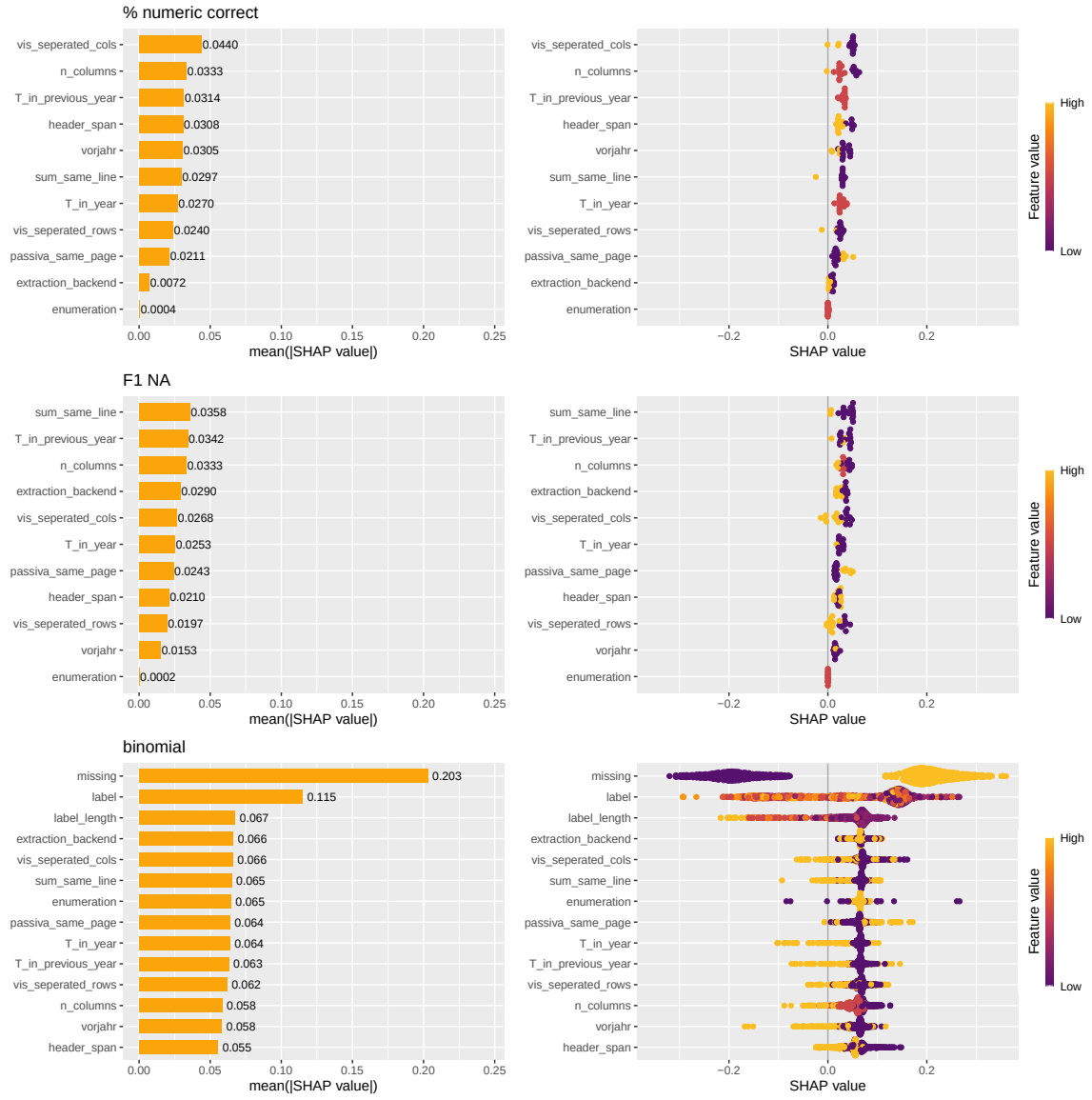


Figure A.1: Mean absolute SHAP values and beeswarm plots for real table extraction with regular expression approach

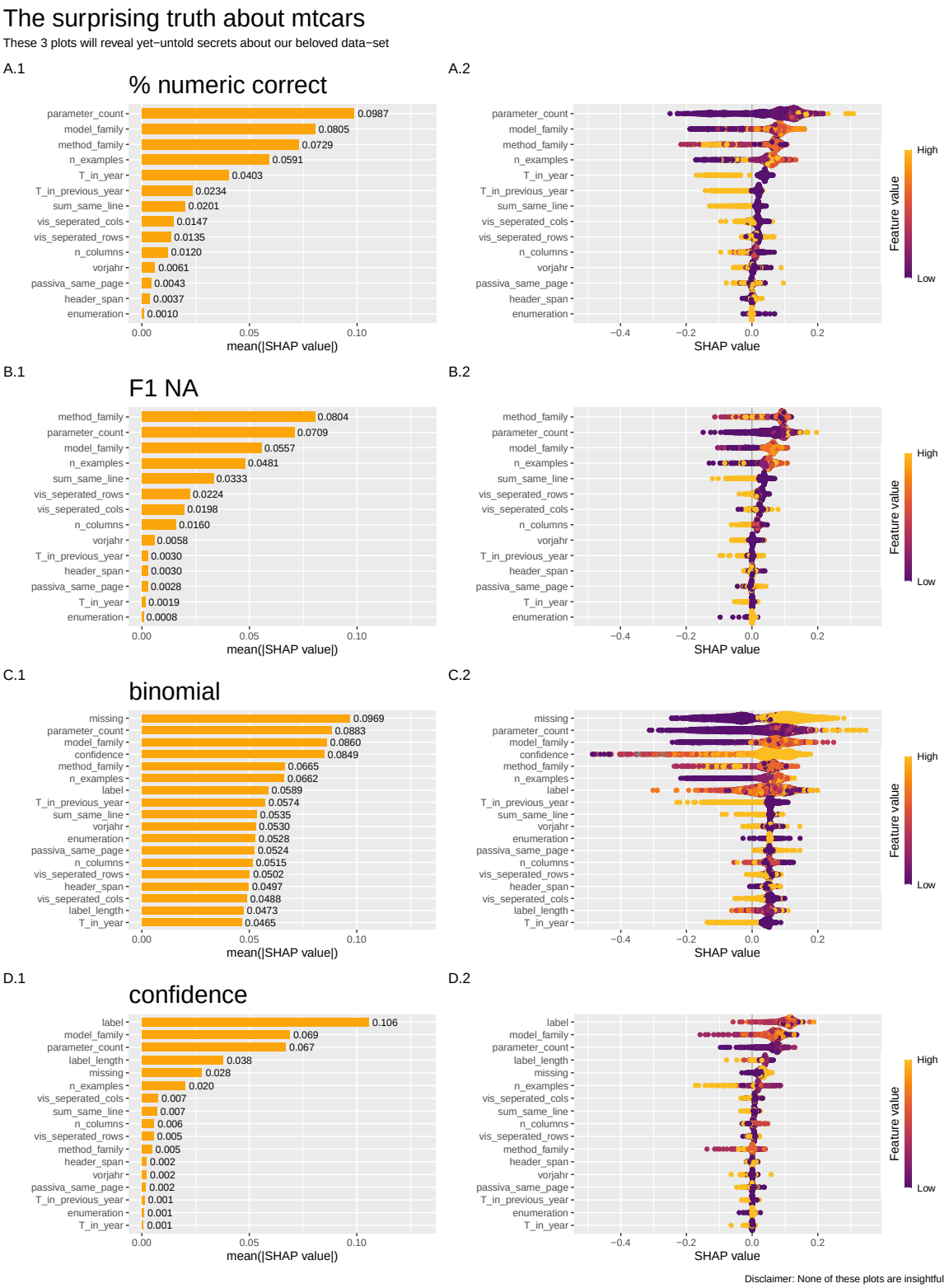


Figure A.2: Mean absolute SHAP values and beeswarm plots for real table extraction with LLMs

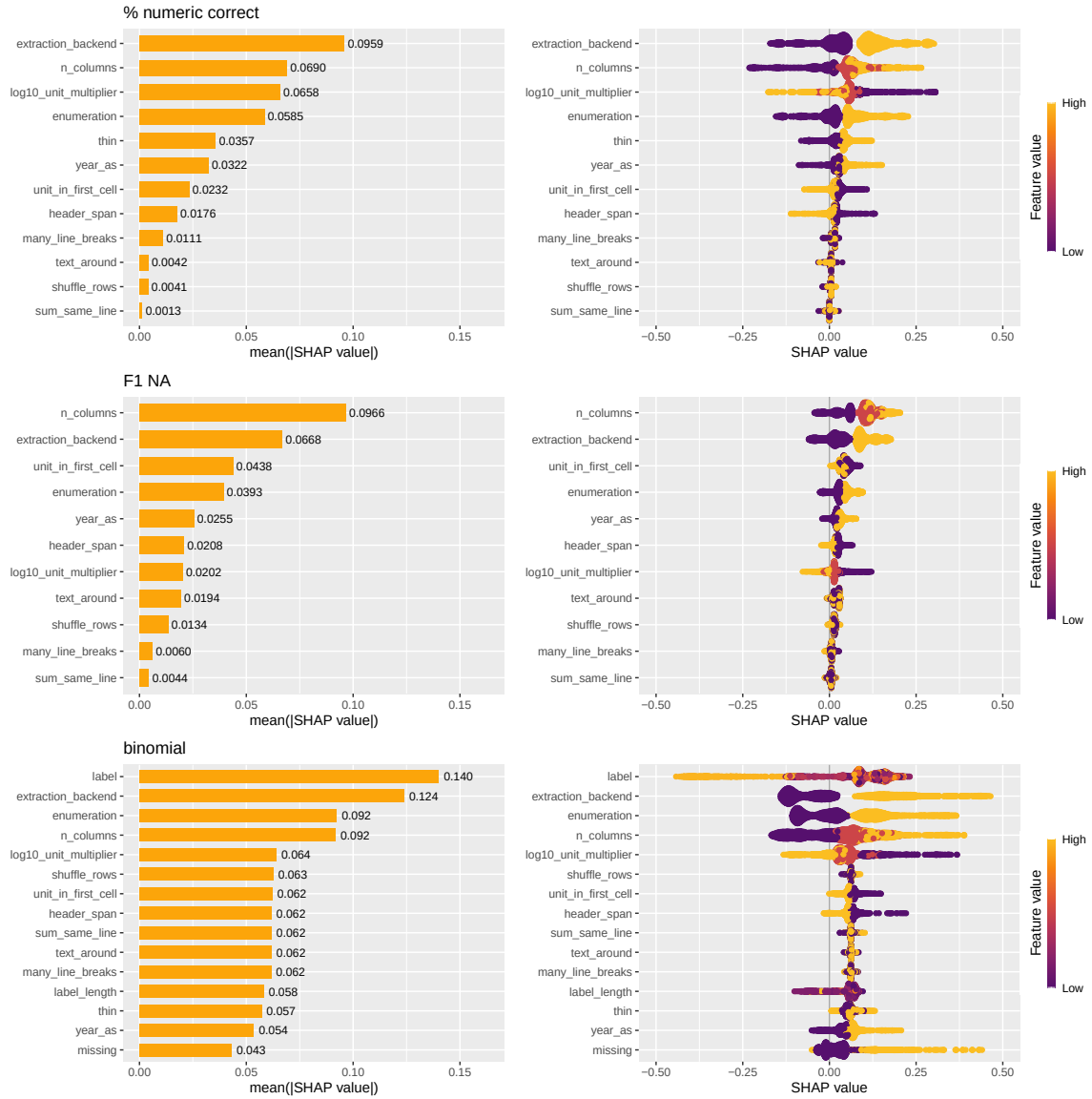


Figure A.3: Mean absolute SHAP values and beeswarm plots for synth table extraction with regular expression approach

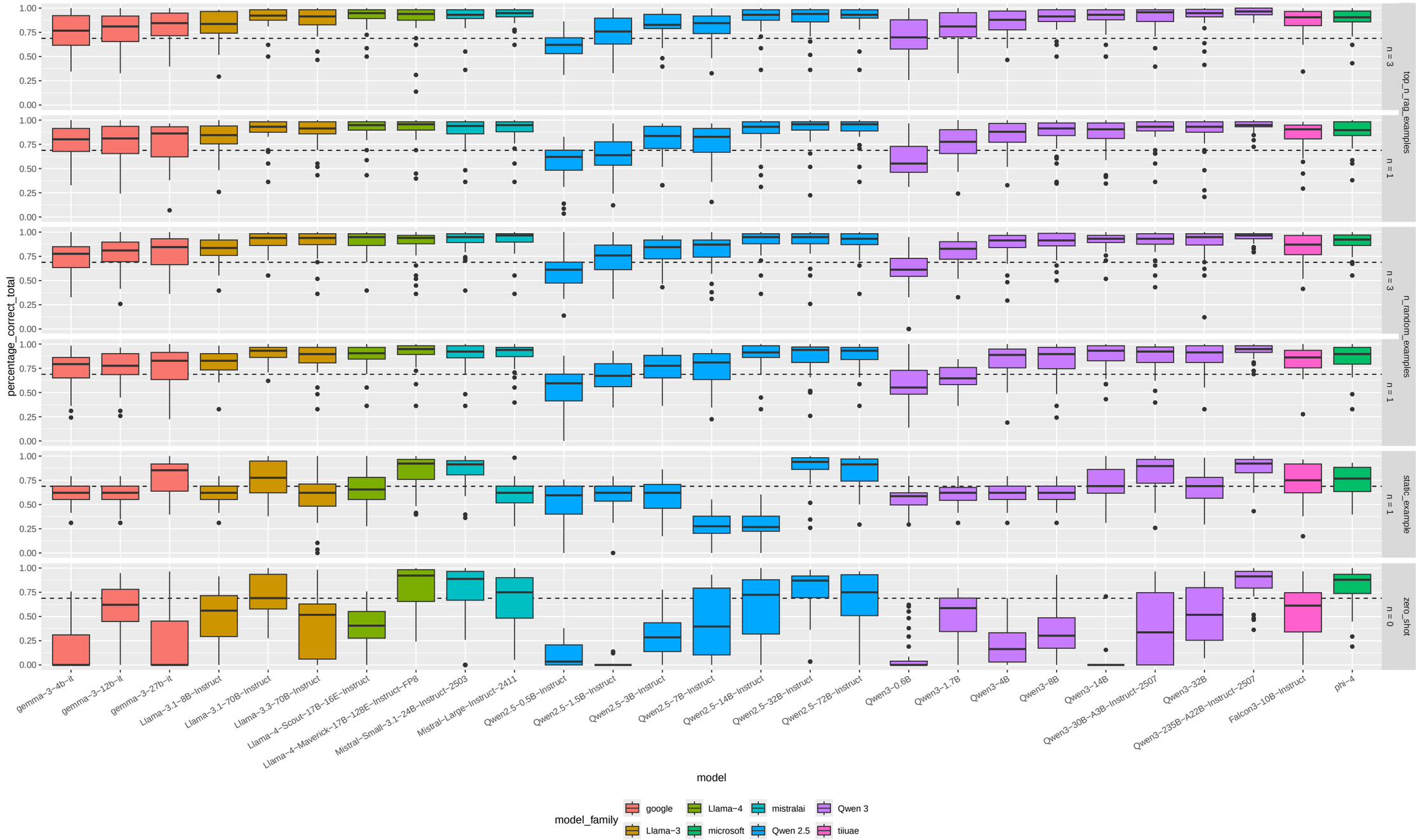


Figure A.4: Percentage of correct extracted or as missing categorized values for table extraction task on real Aktiva tables

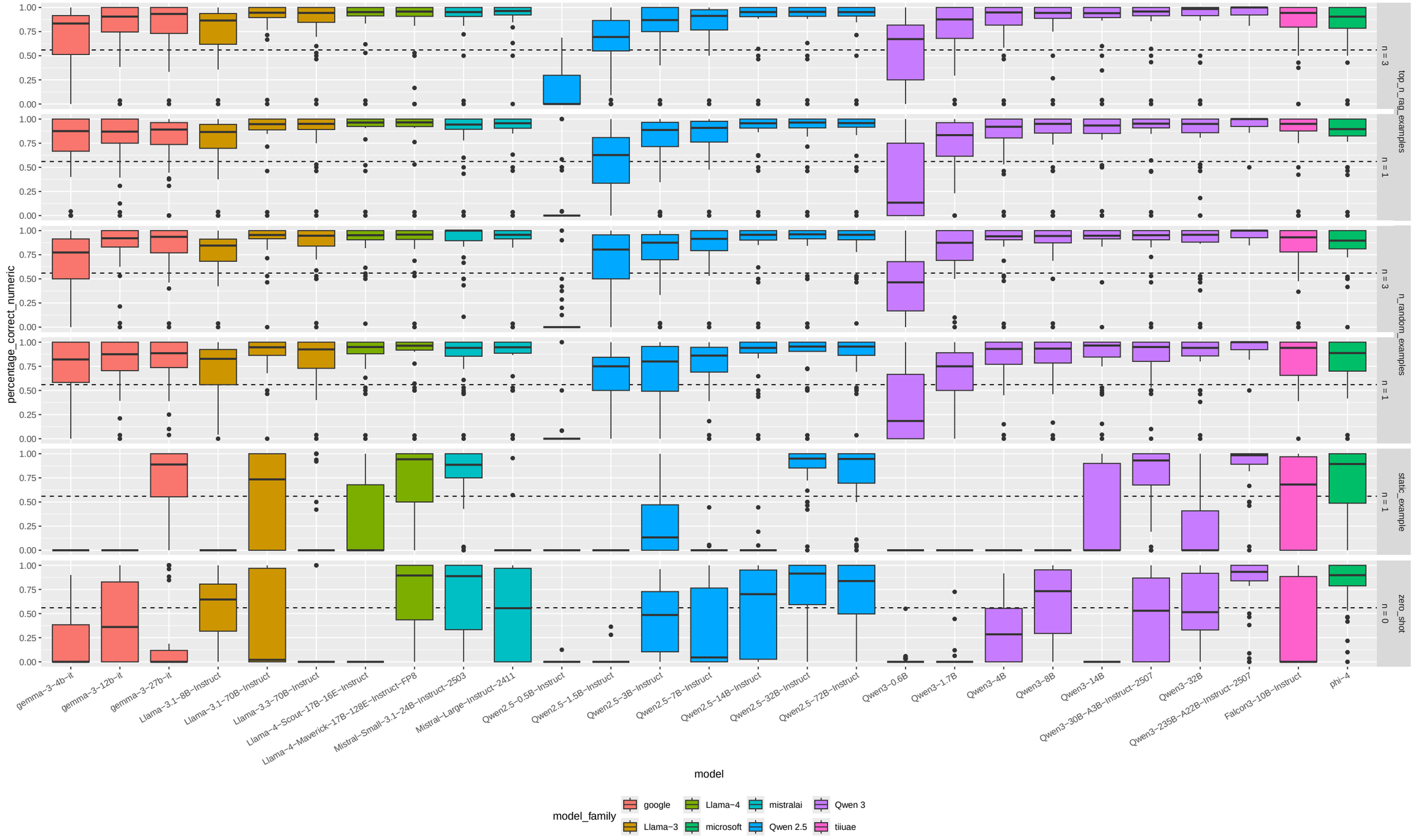


Figure A.5: Percentage of correct extracted numeric values for table extraction task on real Aktiva tables

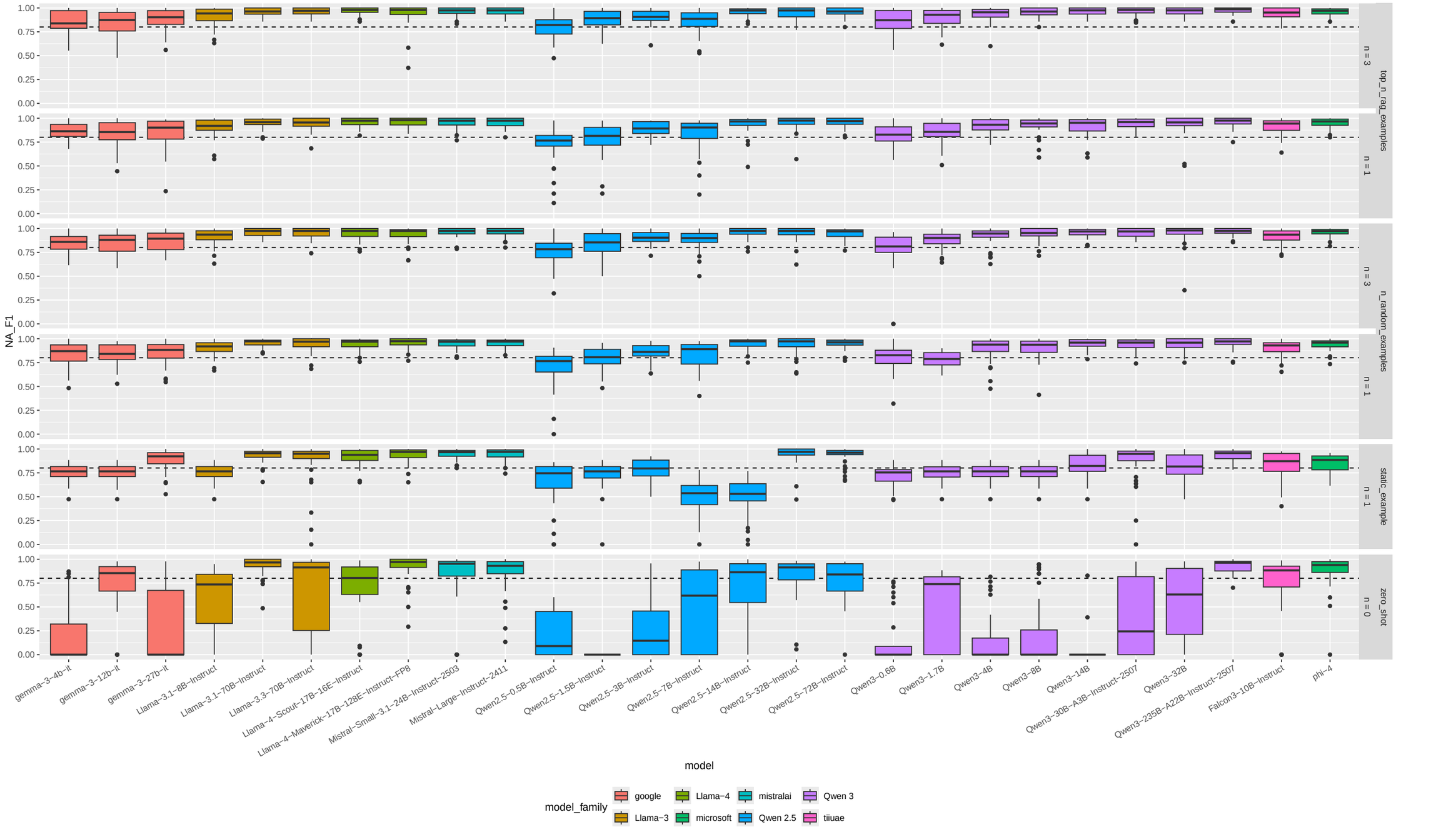


Figure A.6: F1 score for the missing classification if a value is missing for table extraction task on real Aktiva tables

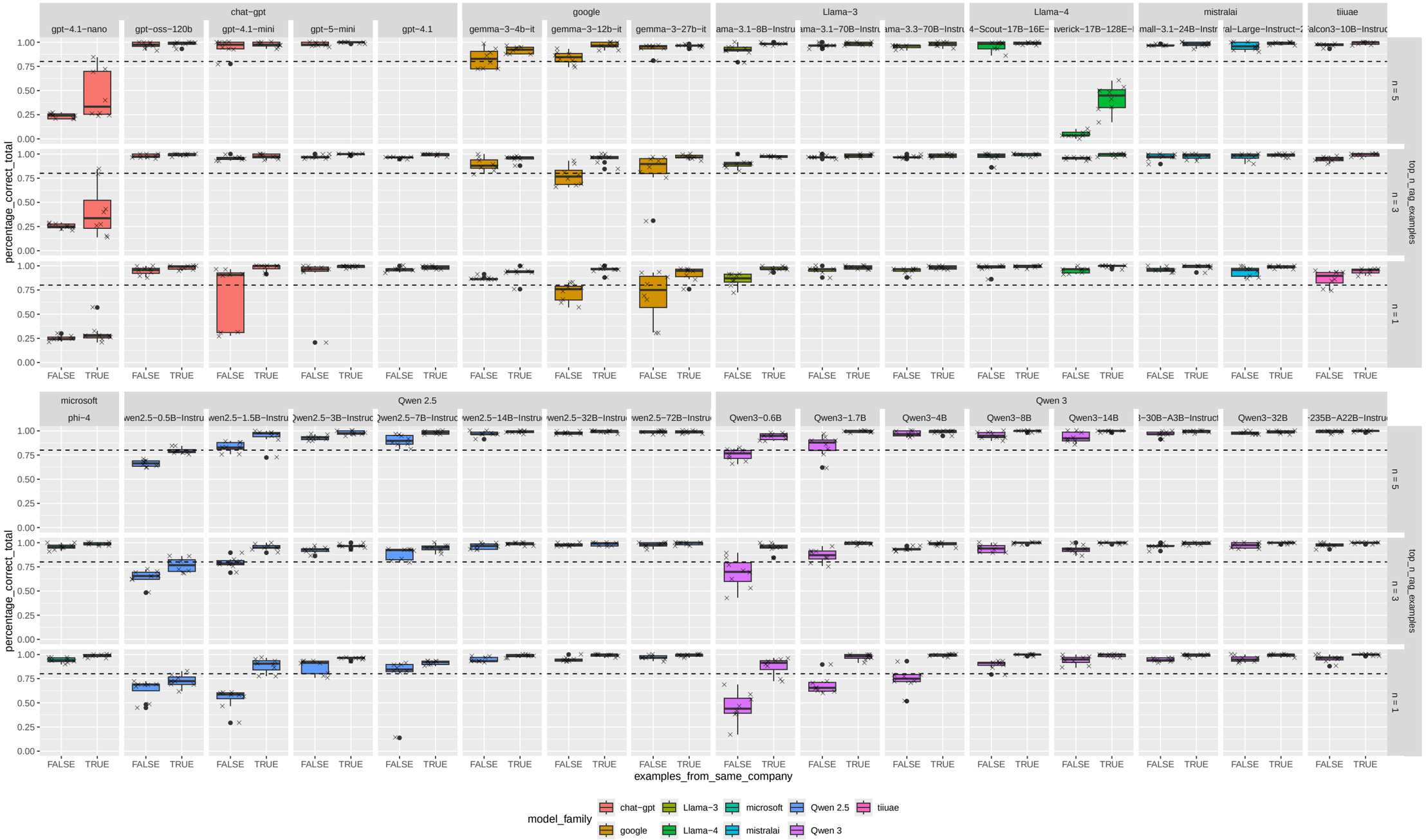


Figure A.7: F1 score for the missing classification if a value is missing for table extraction task on real Aktiva tables

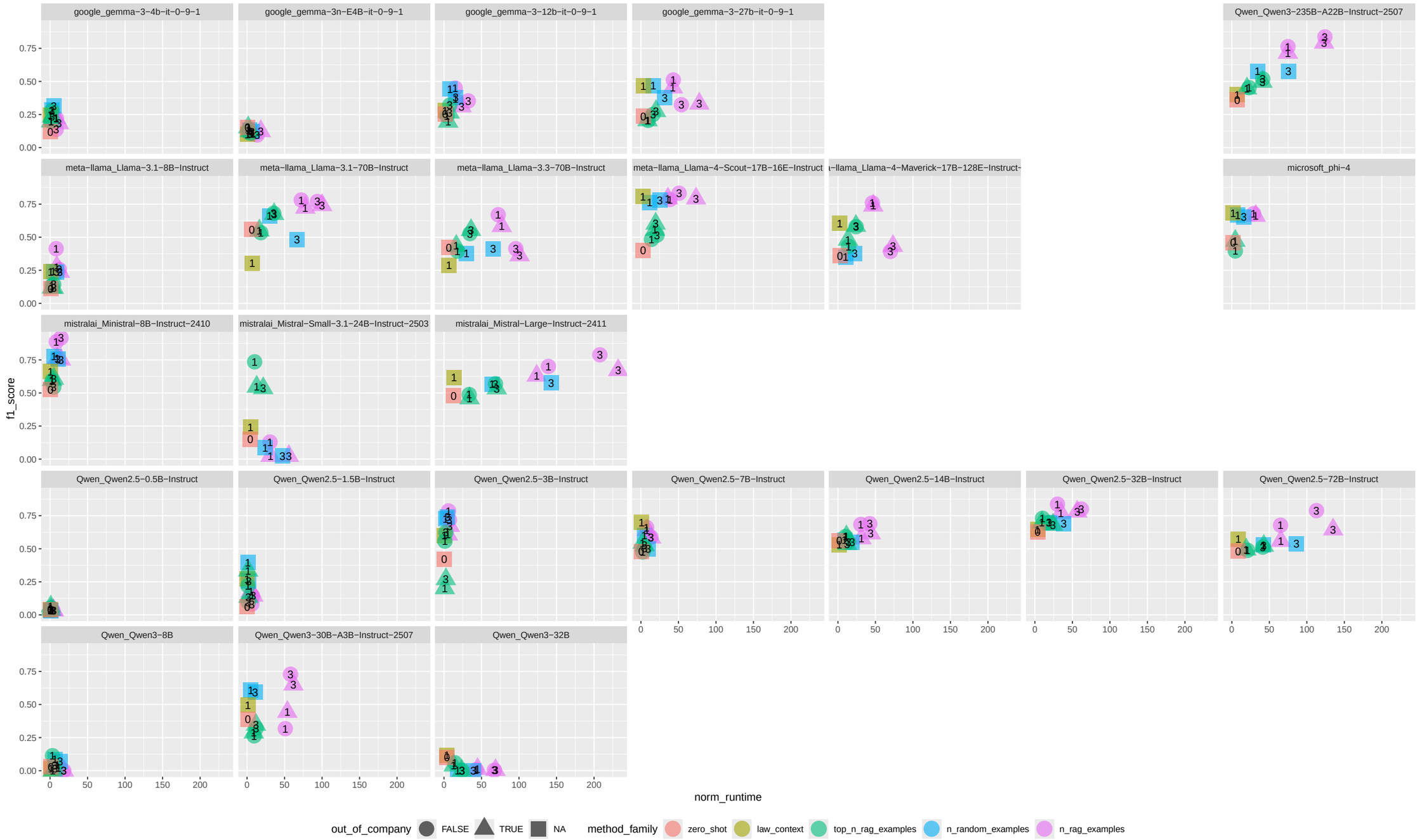


Figure A.8: Comparing F1 score over normalized runtime for binary classification task. The normalized runtime is given in minutes of processing on a single B200. The time to load the model into the VRAM is excluded.

12sdgsdgj

Landscape

A.5 Annual Comprehensive Financial Report Balance Sheet

A.6 Extraction framework flow chart

A.7 Table extraction with regular expressions

Extract by pdfium for '../../benchmark_truth/synthetic_tables/separate_files/final/aktiva_table__3_columns__span_False__thin_Fa
€__enumeration_False__shuffle_True__text_around_True__max_length_50__sum_in_same_row_False__0.pdf':

```
A

ktiva(inMio. €)GeschäftsjahrVorjahr

Anlagevermögen Immaterielle Verm

ögensgegenstände

SelbstgeschaffenegewerblicheSchutzrechteund

ähnlicheRechteundWerte

0,184,77

Geschäfts-oderFirmenwert4,426,78

geleisteteAnzahlungen1,780,65

entgeltlicherworbeneKonzessionen, gewerbliche

SchutzrechteundähnlicheRechteundWertesowie

LizenzenansolchenRechtenundWerten

4,646,71

11,0218,91

Sachanlagen

Grundstücke, grundstücksgleicheRechteundBauten

einschließlichderBautenauffremdenGrundstücken

2,802,55

TechnischeAnlagenundMaschinen5,205,53

AndereAnlagen, Betriebs-undGeschäftsausstattung1,601,93

geleisteteAnzahlungenundAnlagen imBau3,255,81

12,8615,83
```

State of California Annual Comprehensive Financial Report

Balance Sheet

Governmental Funds

June 30, 2023
(amounts in thousands)

	General	Federal
ASSETS		
Cash and pooled investments	$ 71,968,861	$ 6,986,275
Investments	—	—
Receivables (net)	46,621,774	2,076,598
Due from other funds	6,933,803	165,231
Due from other governments	4,075,837	37,069,188
Interfund receivables	3,914,413	—
Loans receivable	45,225	384,293
Other assets	6,244	601,252
Total assets	**$ 133,566,157**	**$ 47,282,837**
LIABILITIES		
Accounts payable	$ 14,422,777	$ 24,499,200
Due to other funds	3,911,973	3,865,533
Due to component units	264,995	—
Due to other governments	21,808,112	11,125,464
Interfund payables	2,692,941	—
Benefits payable	—	69,623
Revenues received in advance	25,891	6,675,956
Tax overpayments	21,740,974	—
Deposits	4,231	—
Unclaimed property liability	1,314,797	—
Other liabilities	522,844	46,256,400
Total liabilities	**66,709,535**	**92,492,176**
DEFERRED INFLOWS OF RESOURCES	2,852,934	10,709
Total liabilities and deferred inflows of resources	**69,562,469**	**92,502,885**
FUND BALANCES		
Nonspendable	3,950,919	—
Restricted	24,830,454	1,210,267
Committed	4,210,891	—
Assigned	20,714,283	—
Unassigned	10,297,141	(46,430,315)
Total fund balances (deficit)	**64,003,688**	**(45,220,048)**
Total liabilities, deferred inflows of resources, and fund balances	**$ 133,566,157**	**$ 47,282,837**

 The notes to the financial statements are an integral part of this statement.

Figure A.9: Example balance sheet pagefom Californias Annual Comprehensive Financial Report 2023

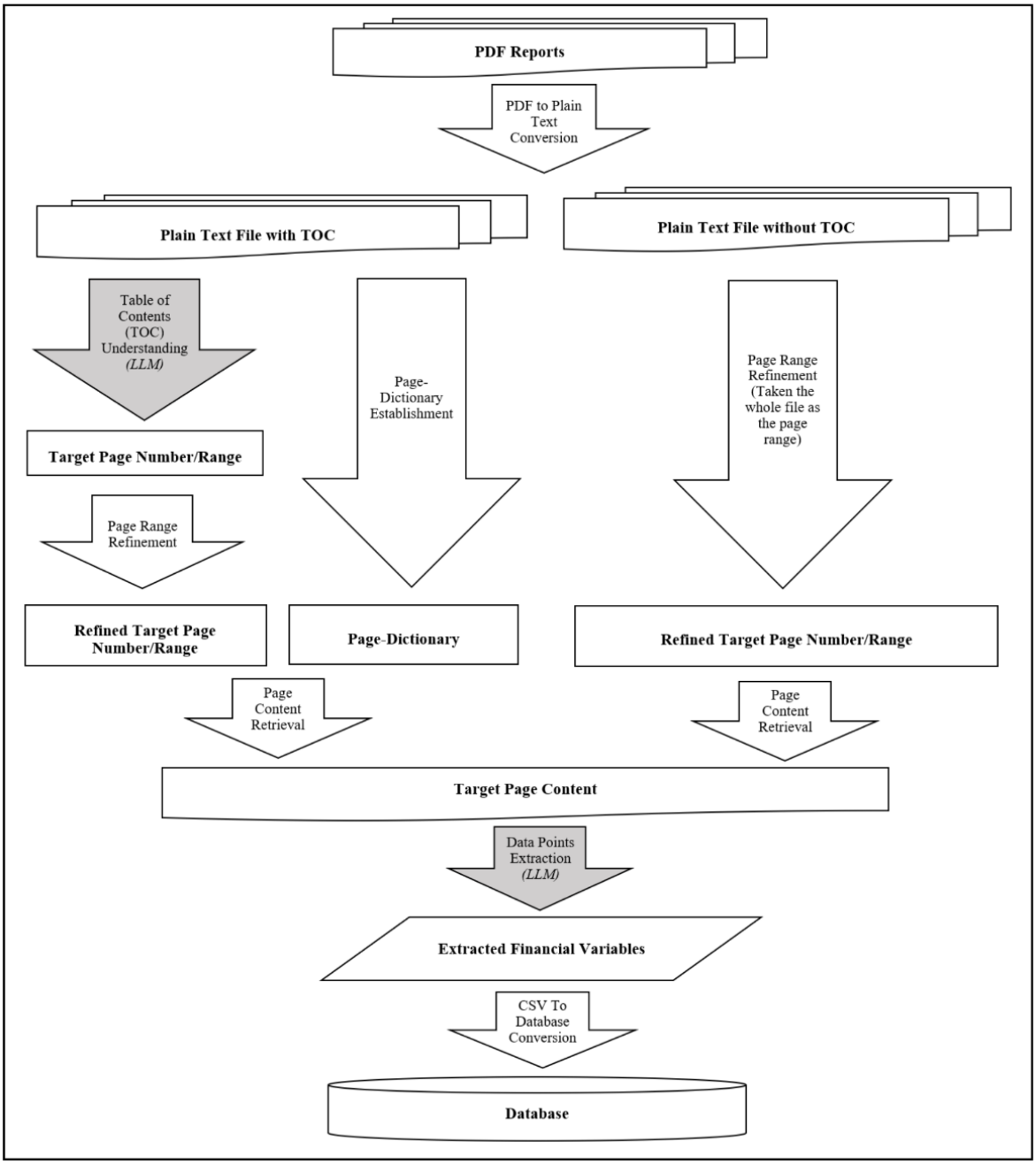


Figure A.10: Flowchart of the extraction framework of Li et al. (2023)

```
Finanzanlagen

SonstigeFinanzanlagen7,446,51

AnteileanverbundenenUnternehmen0,499,83

AusleihungenanverbundeneUnternehmen0,573,49

Beteiligungen1,059,43

AusleihungenanUnternehmen, mitdenenein

Beteiligungsverhältnisbesteht

6,957,65

WertpapieredesAnlagevermögens2,002,71

SonstigeAusleihungen9,091,52

27,5841,13

51,4675,87

Umlaufvermögen

Vorräte

Roh-, Hilfs-undBetriebsstoffe0,382,98

UnfertigeErzeugnisse, unfertigeLeistungen3,236,19

FertigeErzeugnisseundWaren6,724,98

GeleisteteAnzahlungen4,024,83

14,3418,98

ForderungenundsonstigeVermögensgegenstände

ForderungenausLieferungenundLeistungen4,328,36

ForderungengegenverbundeneUnternehmen6,082,38

ForderungengegenUnternehmen, mitdenenein

Beteiligungsverhältnisbesteht

7,878,11

SonstigeVermögensgegenstände1,968,30

20,2227,15
```

```
Wertpapiere

AnteileanverbundenenUnternehmen2,383,24

SonstigeWertpapiere0,077,65

2,4410,88

Kassenbestand, Bundesbankguthaben, Guthabenbei Kreditinstituten und Schecks

4,144,00

41,1561,01

Rechnungsabgrenzungsposten2,746,78

Aktive latenteSteuern8,464,60

AktiverUnterschiedsbetragausder

Vermögensverrechnung

2,863,35

106,67151,61
```

Extract by pdfminer for '../../benchmark_truth/synthetic_tables/separate_files/final/aktiva_table__3_columns__span_False__thin_
€__enumeration_False__shuffle_True__text_around_True__max_length_50__sum_in_same_row_False__0.pdf':

```
Aktiva (in Mio. €)

Anlagevermögen

Immaterielle Vermögensgegenstände

Selbst geschaffene gewerbliche Schutzrechte und
ähnliche Rechte und Werte

Geschäfts- oder Firmenwert

geleistete Anzahlungen

entgeltlich erworbene Konzessionen, gewerbliche
Schutzrechte und ähnliche Rechte und Werte sowie
Lizenzen an solchen Rechten und Werten

Sachanlagen

Grundstücke, grundstücksgleiche Rechte und Bauten
einschließlich der Bauten auf fremden Grundstücken

Technische Anlagen und Maschinen
```

```
Andere Anlagen, Betriebs- und Geschäftsausstattung

geleistete Anzahlungen und Anlagen im Bau

Finanzanlagen

Sonstige Finanzanlagen

Anteile an verbundenen Unternehmen

Ausleihungen an verbundene Unternehmen

Beteiligungen

Ausleihungen an Unternehmen, mit denen ein
Beteiligungsverhältnis besteht

Wertpapiere des Anlagevermögens

Sonstige Ausleihungen

Umlaufvermögen

Vorräte

Roh-, Hilfs- und Betriebsstoffe

Unfertige Erzeugnisse, unfertige Leistungen

Fertige Erzeugnisse und Waren

Geleistete Anzahlungen

Forderungen und sonstige Vermögensgegenstände

Forderungen aus Lieferungen und Leistungen

Forderungen gegen verbundene Unternehmen

Forderungen gegen Unternehmen, mit denen ein
Beteiligungsverhältnis besteht

Sonstige Vermögensgegenstände

Wertpapiere

Anteile an verbundenen Unternehmen

Sonstige Wertpapiere

Kassenbestand, Bundesbankguthaben, Guthaben bei
Kreditinstituten und Schecks

Rechnungsabgrenzungsposten
```

```
Aktive latente Steuern

Aktiver Unterschiedsbetrag aus der
Vermögensverrechnung

Geschäftsjahr

Vorjahr

0,18

4,42

1,78

4,64

11,02

2,80

5,20

1,60

3,25

12,86

7,44

0,49

0,57

1,05

6,95

2,00

9,09

27,58

51,46

0,38

3,23

6,72
```

```
4,02

14,34

4,32

6,08

7,87

1,96

20,22

2,38

0,07

2,44

4,14

41,15

2,74

8,46

2,86

4,77

6,78

0,65

6,71

18,91

2,55

5,53

1,93

5,81

15,83

6,51

9,83
```

```
3,49

9,43

7,65

2,71

1,52

41,13

75,87

2,98

6,19

4,98

4,83

18,98

8,36

2,38

8,11

8,30

27,15

3,24

7,65

10,88

4,00

61,01

6,78

4,60

3,35

106,67

151,61
```