End-to-end Information Extraction from Archival Records with Multimodal Large Language Models

Mahsa Vafaie
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology
Karlsruhe, Germany
mahsa.vafaie@fiz-karlsruhe.de

Sven Hertling
Data and Web Science Group
University of Mannheim
Mannheim, Germany
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Eggenstein-Leopoldshafen, Germany
sven.hertling@fiz-karlsruhe.de

Inger Banse-Strobel
Bundesarchiv
Koblenz, Germany
inger.banse-strobel@bundesarchiv.de

Kevin Dubout
Bundesarchiv
Koblenz, Germany
k.dubout@bundesarchiv.de

Harald Sack
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Eggenstein-Leopoldshafen, Germany
Karlsruhe Institute of Technology
Karlsruhe, Germany
harald.sack@fiz-karlsruhe.de

Abstract

Semi-structured Document Understanding presents a challenging research task due to the significant variations in layout, style, font, and content of documents. This complexity is further amplified when dealing with *born-analogue* historical documents, such as digitised archival records, which contain degraded print, handwritten annotations, stamps, marginalia and inconsistent formatting resulting from historical production and digitisation processes. Traditional approaches for extracting information from semi-structured documents rely on manual labour, making them costly and inefficient. This is partly due to the fact that within document collections, there are various layout types, each requiring customised optimisation to account for structural differences, which substantially increases the effort needed to achieve consistent quality. The emergence of Multimodal Large Language Models (MLLMs) has significantly advanced Document Understanding by enabling flexible, prompt-based understanding of document images, needless of OCR outputs or layout encodings. Moreover, the encoder-decoder architectures have overcome the limitations of encoder-only models, such as reliance on annotated datasets and fixed input lengths. However, there still remains a gap in effectively applying these models in real-world scenarios. To address this gap, we first introduce BZKOpen, a new annotated dataset designed for key information extraction from historical German index cards. Furthermore, we systematically assess the capabilities of several state-of-the-art MLLMs—including the open-source InternVL2.0 and InternVL2.5 series, and the commercial GPT-4o-mini— on the task of extracting key information from these archival documents. Both zero-shot and few-shot prompting strategies are evaluated across different model configurations to identify the optimal conditions for performance. Interestingly, our results reveal that increasing model size does not necessarily lead to better performance on this dataset. Among all models tested, the open-source InternVL2.5-38B consistently achieves the most robust results, outperforming both larger InternVL models and the proprietary alternative. We further provide practical insights into prompt engineering and inference settings, offering guidance for applying MLLMs to real-world key information extraction tasks. Additionally, we highlight the need for more ground truth datasets that include a wider range of historical documents with varying quality and in multiple languages, in order to fully explore the potentials and limitations of MLLMs for key information extraction from historical records.

CCS Concepts

- **Information systems → Information extraction**.

Keywords

Multimodal Large Language Models, Document Understanding, Key Information Extraction, Digital Cultural Heritage

ACM Reference Format:
Mahsa Vafaie, Sven Hertling, Inger Banse-Strobel, Kevin Dubout, and Harald Sack. 2025. End-to-end Information Extraction from Archival Records with Multimodal Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746252.3761503

1 Introduction

Ensuring equitable public access to archival records is increasingly recognised as a matter of social justice [10, 32]. Archives are not only repositories of institutional memory but also vital instruments

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761503

for community identity, historical accountability, and civic participation. As Duff and Haskell [9] argues, archival systems must be reimagined to support inclusive and flexible access frameworks that empower individuals and communities to engage meaningfully with historical records. Digitalisation offers a critical pathway to this transformation, enabling the preservation and dissemination of materials that have long remained hard to access. Digitalisation pipelines employing a combination of nascent computer vision, natural language processing, artificial intelligence and semantic web technologies offer a way forward with broadening public access to archival materials stored in archival institutions as part of the shared cultural heritage, and promote novel research across disciplines [28].

Within this broader vision of democratised, digital archival access, the field of Document Understanding plays a pivotal role. Central to this field is the task of Key Information Extraction (KIE)—the automated identification and structuring of meaningful data in the form of key-value pairs from diverse and often visually complex document formats. Despite recent advances, KIE continues to pose significant challenges to digitalisation pipelines [31]. In the context of born-analogue archival records, such as historical index cards, this task becomes especially demanding [27]. Historical archival documents often lack standardised layouts and may include a mix of printed text, handwritten notes, stamps, and annotations that challenge off-the-shelf Optical Character Recognition (OCR) models and conventional information extraction pipelines [39]. Accurate extraction of information in such scenarios demands a deep understanding of contextual semantics, precise spatial alignment among textual components, and the ability to interpret hierarchical entities arranged in non-linear formats. Advances in Multimodal Large Language Models (MLLMs) and document understanding transformers have opened new avenues for tackling the task of KIE from these multimodal, noisy inputs [7]. By exploring the effectiveness of these models in a real-world archival scenario, we aim to contribute to more inclusive and accessible infrastructures for historical research and cultural heritage representation.

In this paper, we first introduce a novel dataset designed to support research on key information extraction and then investigate the use of MLLMs for automated information extraction from historical archival materials using this dataset, focusing on German index cards as a representative case study. The study draws on a card file comprising about 1.9 million digitised index cards from the State Offices for Compensation (*Ämter für Wiedergutmachung*[1] in German), which document claims and proceedings related to material compensation for persecutees of National Socialism in the Federal Republic of Germany. These records form part of a broader corpus spanning over 100 km of archival material, which will be integrated into the Online Collection *"Themenportal Wiedergutmachung nationalsozialistischen Unrechts"*[2].

The diversity and complexity of these records present significant challenges for traditional methods of KIE, making them a valuable testbed for evaluating the performance of the more recent transformer-based models and MLLMs in real-world archival scenarios. We suggest that the proposed MLLM-based KIE pipeline developed for this card file not only enables semantic enrichment and access within this specific use case, but also provides a scalable and adaptable framework that can be applied to other large-scale, born-analogue archival collections with similar characteristics, making it a transferable model for improving access to memory institutions and promoting inclusive archival research.

Following a review of relevant literature, different methods, and various datasets for information extraction from semi-structured documents, this paper describes a new dataset designed for information extraction from historical archival documents, as well as the designed pipeline for end-to-end information extraction from such records. After presenting the conducted experiments and their results, the key findings of our studies are discussed and a critical reflection on our approach is presented. Finally, the paper is concluded by sketching the deployment plans and next steps for further assessing, improving and building upon the presented pipeline.

2 Related Work

Document Key Information Extraction (KIE), a specialised task within the field of Document Understanding, focuses on identifying and extracting a set of pre-defined entities from semi-structured documents like invoices, receipts, and forms. This task has emerged as a central focus in both industrial applications and academic research, driven by its pivotal role in enabling intelligent document processing. Consequently, a diverse body of work has been dedicated to advancing KIE techniques across heterogeneous document formats. In this section, we first review the principal methods developed for KIE, ranging from rule-based systems to recent transformer-based approaches, and briefly describe existing datasets commonly used to evaluate performance across diverse document types.

2.1 Document Key Information Extraction methods

Early methods for document KIE relied on manually crafted rules and regular expressions to extract entities from structured and semi-structured documents, but these approaches struggle with complex or varying layouts [11, 13]. The development of neural models made it possible to combine different modalities, resulting in more robust and generalised information extraction. Deep learning-based approaches to KIE typically fall into three categories: sequence-based models, layout-aware models, and pre-trained multimodal transformers. Sequence-based models treat token sequences linearly and can capture contextual dependencies, but often neglect layout structure [16, 24].

A major advancement has been the introduction of layout-aware pre-trained models such as LayoutLM [14], SelfDoc [23], and StrucText [25], which embed positional, textual, and visual information into unified representations. These models have demonstrated superior performance by leveraging large-scale pretraining on document-rich corpora [26, 45, 46].

Despite these advances, KIE remains challenging due to variability in document templates, noisy OCR outputs, and the scarcity of annotated datasets. Consequently, few-shot and zero-shot learning strategies are gaining attention, as they aim to reduce dependence

[1] "Wiedergutmachung" is a German word that translates to "making good again" or "making amends". In the context of National Socialism in Germany and its aftermath, it specifically refers to the efforts made to compensate survivors and persecutees for the losses they suffered during the rule of the Nazi regime.

[2] https://www.archivportal-d.de/themenportale/wiedergutmachung

on labeled data and OCR outputs while maintaining extraction accuracy. However, there remains a gap in the literature in evaluating general-purpose MLLMs for KIE from real-word document understanding scenarios.

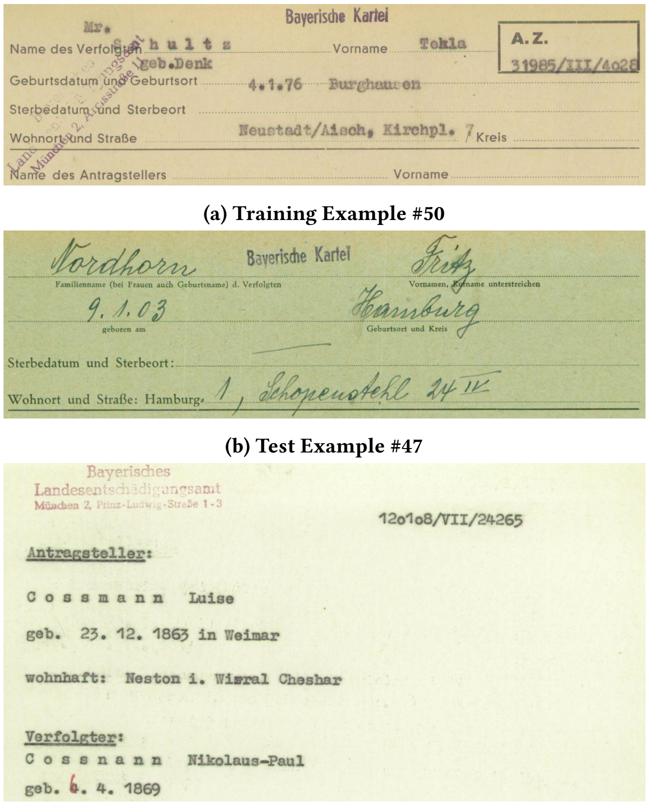
2.2 Datasets for Key Information Extraction from semi-structured Documents

Several benchmark datasets have been widely used for KIE from semi-structured documents, each designed to represent distinct real-world challenges. FUNSD [17] (Form Understanding in Noisy Scanned Documents) comprises 199 scanned forms with manual annotations of keys, values, and semantic relationships, targeting the understanding of complex, noisy forms that are common in administrative workflows. To address the extraction of information from receipts, the SROIE dataset [15] was introduced as part of the ICDAR 2019 competition [34]; it features 1,000 scanned receipts annotated with four key-value pairs, reflecting real-world retail scenarios. For more complex, multi-page legal and administrative documents, the Kleister-NDA dataset [36] was developed, focusing on non-disclosure agreements annotated for a range of fields pertinent to contract analysis. In addition, datasets such as CORD (Consolidated Receipt Dataset) [30] extend KIE research to non-English and multilingual settings, providing Korean receipts annotated for key-value pairs, while DocBank [22] offers automatic large-scale annotation of PDF scientific articles for comprehensive document layout understanding. These datasets have collectively enabled rigorous and reproducible evaluation of KIE systems across a broad spectrum of real-world document types. Nonetheless, existing datasets do not adequately reflect the unique characteristics of historical archival materials, such as card files. To address this gap, we introduce a novel dataset specifically designed to support the evaluation of KIE methods on this underrepresented document type.

3 BZKOpen Dataset

Despite ongoing efforts in the creation of datasets for information extraction from semi-structured documents, to the best of our knowledge, there is currently no publicly available dataset focused on KIE from historical documents in German that also encompasses a wide variety of layout types and complex structural features. To address this gap, we introduce BZKOpen, a novel German-language dataset designed specifically for KIE, comprising annotations for 19 attributes across 516 historical index cards, appearing on more than 40 layout types.

The documents are sourced from a card file called *Bundeszentralkartei (BZK)* (described in detail below), with careful selection to comply with privacy constraints while ensuring a diversity of layout styles. As illustrated in Figure 1, the dataset presents a range of real-world complexities including overlapping stamps, handwritten entries, spatially misaligned key-value fields, and diverse document layouts. To support reproducible research and further advancements in historical document understanding, and to facilitate robust model training and evaluation, the dataset is partitioned into 70/15/15 splits for training, validation, and testing, respectively, and released for public use on Hugging Face[3].



Bayerische Kartei

Mr.

Name des Verfolgten Schultz Vorname Tekla

geb.Denk

A.Z.

31985/III/4028

Geburtsdatum und Geburtsort 4.1.76 Burghausen

Sterbedatum und Sterbeort

Wohnort und Straße Neustadt/Aisch, Kirchpl. 7 Kreis

Name des Antragstellers Vorname

Land[illegible]ungsamt

München 2, Arcisstraße 11

(a) Training Example #50

Nordhorn Bayerische Kartei Fritz

Familienname (bei Frauen auch Geburtsname) d. Verfolgten Vornamen, Rufname unterstreichen

9.1.03 Hamburg

geboren am Geburtsort und Kreis

Sterbedatum und Sterbeort:

Wohnort und Straße: Hamburg 1, Schopenstehl 24 IV

(b) Test Example #47

Bayerisches

Landesentschädigungsamt

München 2, Prinz-Ludwig-Straße 1-3

12o1o8/VII/24265

Antragsteller:

Cossmann Luise

geb. 23. 12. 1863 in Weimar

wohnhaft: Neston i. Wisral Cheshar

Verfolgter:

Cossnann Nikolaus-Paul

geb. 6. 4. 1869

(c) Test Example #9

Figure 1: Crops from three sample documents demonstrating the challenges of the dataset with distinct layouts, overlapping handwritten text and stamps (1a, 1b) and missing keys (1c) Source: Landesarchiv NRW Abteilung Rheinland, BR 3015, BZK-Nr. 31985/III/4028, 11026N7, 120108/VII/24265

The Bundeszentralkartei (BZK) Index Cards. The *Bundeszentralkartei (BZK)* is the central registry of all applications for compensation made in the Federal Republic of Germany, mainly under the *Bundesentschädigungsgesetz (BEG)*, or the Federal Compensation Act. First enacted in 1953, the *BEG* was a crucial part of *Wiedergutmachung* and the BZK was created in its immediate aftermath. Currently, the BZK card file contains approximately 1.9 million index cards, each recording essential details related to compensation claims. These details typically include the names of applicants and persecutees, their dates of birth if available, the applicants' address at the time of application, case reference numbers, and the responsible compensation authority. As a central information resource, the BZK serves to identify whether an individual has submitted a compensation application (or if an application was submitted for them) and, if so, to provide the corresponding reference number and the specific compensation authority responsible for handling the case.

Annotation. To enable a comprehensive evaluation of model performance with our dataset, domain experts from **Bundesarchiv**

[3] https://huggingface.co/datasets/MahsaVafaie/BZKopen

(The Federal Archives of Germany) manually annotated the key-value pairs present in the document images. These ground truth annotations provide the reference framework for assessing the accuracy of key-value extraction under various experimental settings. The annotation process was carried out by three independent annotators, covering 19 distinct fields related to the card itself, the applicant and the persecutee. BZKOpen provides two levels of annotation: raw annotations, which retain the original appearance of key-value pairs as they appear on the documents, and normalised annotations, in which values such as dates are formatted according to the ISO 8601 standard and addresses are simplified to city names.

4 Methodology

For the task of KIE from historical documents with heterogeneous visual and textual layouts, we implemented two pipelines on the BZKOpen dataset. The first pipeline serves as a rule-based baseline, relying on OCR outputs and manually defined regular expressions to detect recurring keys and extract their corresponding values. The second pipeline leverages transformer-based models, with a focus on MLLMs, which offer robust generalisation capabilities across heterogeneous document layouts and allow for end-to-end information extraction from visual inputs, thereby eliminating the need for intermediate OCR transcripts [40].

Following information extraction, all outputs are post-processed and converted into a standardised JSON format, representing the extracted data as key-value pairs to allow for consistent evaluation against ground truth annotations. For enhanced semantic representation, the extracted entities are subsequently transformed into subject–predicate–object triples at the final stage of the pipeline, facilitating integration into a knowledge graph.

4.1 Rule-Based Baseline: Information Extraction with Apache UIMA Ruta

As a baseline for KIE from BZKOpen dataset, we implemented a rule-based pipeline that crucially depends on the availability of accurate OCR transcripts. Transcripts are provided by implementing the TrOCR model TextTitan from Transkribus [4], for its ability to handle documents with a mix of machine-printed and handwritten text. Using Apache UIMA Ruta [19], this approach leverages regular expressions to define patterns for detecting and annotating key information within texts produced by OCR. Post-processing of OCR-generated transcripts is a vital step, addressing common issues such as misspellings, noise, and inconsistent formatting before the main extraction process. The success of the rule-based approach is therefore directly dependent on the quality and consistency of the OCR output. This reliance on high-quality OCR transcripts and substantial manual rule engineering underlines the limitations of the rule-based approach in this context.

4.2 Transformer-based Models

Transformer-based models encompass state-of-the-art architectures capable of learning complex patterns from data with minimal manual intervention. Unlike traditional rule-based approaches, which rely on handcrafted rules, transformer-based methods offer greater flexibility and scalability. Both Donut [18] and Multimodal Large Language Models (MLLMs) belong to this family, leveraging deep learning techniques to jointly process textual and visual information directly from raw document images.

4.2.1 Model Selection. For conducting and evaluating the task of end-to-end KIE from historical documents with the new transformer-based technologies, we utilised the classic Donut framework, as well as multiple open-source and proprietary MLLMs, with systematic prompt engineering and under different prompting strategies.

Donut (Document Understsnding Transformer). Donut [18] is one of the first vision-language models designed for end-to-end document understanding without reliance on OCR. It directly processes document images and generates structured text outputs using an encoder-decoder transformer architecture. Donut formulates key information extraction as a sequence-to-sequence task, allowing joint modeling of layout and semantics. We employed Donut [5] and a variant of Donut fine-tuned on CORD dataset [6], to assess transferability to our own dataset of historical index cards.

MLLMs. Multimodal Large Language Models (MLLMs) are typically built upon transformer architectures, extending them to handle inputs from multiple modalities such as text and images. These models integrate visual and textual information through specialised attention mechanisms and are thus considered a subclass of transformer-based architectures. In this study we selected our models based on the OpenVLM leaderboard[7] which relies on the VLMEvalKit [8], an open-source evaluation toolkit for MLLMs. We found the InternVL series ranking high in this leaderboard (comprising models with various sizes from InternVL2.0 and InternVL2.5) [2–4, 42], and used GPT-4o-mini-2024-07-18 [29] as a commercial counterpart for comparison. GPT-4o-mini serves as a general-purpose commercial baseline, while InternVL models are open-source and freely available. We note that the use of commercial models is not permissible for our use case in deployment phase, as we work with sensitive data subject to strict data privacy protections.

4.2.2 Prompt Engineering. In parallel with the advancement of LLMs to handle complex problems, prompt engineering has gained prominence as a significant method for leveraging these models effectively across diverse tasks and domains [35]. Carefully crafted prompts enable LLMs and MLLMs towards more precise execution of designated tasks.

Iterative Context-Aware Prompting. In this study, the prompt engineering step involves analysing the output of the initial basic prompt (prompt number 0) and using this analysis to refine the prompt further, creating a feedback loop that continually guides the model and enhances the model's performance. At every step of the way, the results of the previous prompts are analysed, and the prompt is modified with some additional hints based on the errors from the previous step. In our experiments we implemented 5 different prompts to get to the optimal one (Prompt number 9), which includes keywords and hints that guide the model for more

[4] https://www.transkribus.org/

[5] https://huggingface.co/naver-clova-ix/donut-base

[6] https://huggingface.co/naver-clova-ix/donut-base-finetuned-cord-v2

[7] https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

accurate extraction of information from the documents (such as hints for German words used for applicant and persecutee in the different layout types). It is worth mentioning that even though our documents are in German, the English language prompts were uniformly performing better than their German translations.

zero-shot and few-shot prompting. Zero-shot and few-shot prompting are two widely adopted strategies for leveraging the generalisation capability of large (multimodal) language models [1, 43]. In the zero-shot paradigm, the model is provided only with task instructions and no task-specific examples, requiring it to rely entirely on its pre-trained knowledge and reasoning abilities. In contrast, few-shot prompting involves enriching the prompt with a small number of curated input-output examples that illustrate the desired task and output format. This approach offers a flexible means to quickly adapt LLMs to new or heterogeneous information extraction settings, as it reduces the need for extensive fine-tuning. We further distinguish between two few-shot strategies: static and dynamic. In the static few-shot approach, the same set of examples is used for all test samples, irrespective of individual input variation. Conversely, the dynamic few-shot strategy dynamically selects support examples from the training set that are most similar to the current test sample [12].

In this study, all three prompting strategies are evaluated to assess their effectiveness on documents with varying layouts, highlighting the potential to generalise across diverse document types using only a handful of representative examples or, in the case of zero-shot, through task specification alone. For similar few shot prompting image similarity is determined by embedding the images with CLIP[8] [33] and using cosine distance.

4.2.3 Output Processing. The output generated by MLLMs is often not directly in a parseable JSON format, requiring several post-processing steps. We first use a more forgiving JSON parser that extracts any content between curly brackets. Within a JSON object, the textual values might not always be surrounded by quotation marks, and thus we also use newlines (in case they are not correctly used within a JSON string) to find the end of that value. Similarly, we used a colon to find the end of a JSON key in case it is not a valid string. The output of the previous step is a parsed key-value data structure, but model-generated keys may deviate from the expected schema (e.g., “ApplicantFirstName” vs. “Applicant_First_Name”). To address this, we compute the edit distance between generated and expected keys and use a greedy approach to align them, prioritising the closest matches. This process standardises the output into a key-value structure with predefined keys, enabling reliable access to the extracted information.

4.3 Knowledge Graph Generation

Knowledge graphs are essential for linking archival records to the Semantic Web, transforming isolated data points into interconnected nodes. By interlinking extracted information, knowledge graphs allow researchers to explore complex relationships and uncover patterns that extend beyond individual documents [41]. To achieve this, the structured data in JSON format containing information from the archival records is converted into RDF triples using the CourtDocs ontology [38] and dedicated data models developed for this use case [37]. The resulting knowledge graph will serve as a core component of the online collection *“Themenportal Wiedergutmachung nationalsozialistischen Unrechts”* [9] ”.

5 Experiments & Results

The primary research question addressed in this study concerns the effectiveness of transformer-based models for KIE from historical semi-structured documents. To rigorously evaluate these models, it is essential to establish a baseline for comparison. For this purpose, we implemented a rule-based approach on a subset of our dataset comprising 75 index cards from three distinct layout classes, each representing cards from compensation authorities in Berlin, Baden-Württemberg (BW), and Bayern. While this test set does not fully overlap with the test split of BZKOpen, we selected layout classes such that the sample size closely matches our standard test set of 78 cards. A key motivation for employing transformer models in the context of this study is to reduce the need for manual labour and handcrafted information extraction rules; following this rationale, we limited the rule-based baseline to three layout types, as creating customised rules for all layout types would be extremely time-consuming. The results of the rule-based baseline are presented in Table 1.

Model	Layout Type	No. of Cards	EM (t=0)	PM (t=1)
Rule-based	Berlin	38	82%	88%
	BW	15	79%	84%
	Bayern	22	62%	69%
	Average	75	74%	80%

Table 1: Performance of the rule-based approach across three different layout types and their average from a subset of the BZKOpen dataset with 75 cards.

Then, to compare the specialised document understanding transformer (Donut) with the general transformer-based MLLMs—both open-source and commercial models, and across various sizes—we conducted multiple experiments using both their pre-trained versions and versions fine-tuned on our train set, which consists of 361 cards and their ground truth annotations(Table 2).

We further evaluated the impact of prompting strategies on the top-performing InternVL2.5-38B model and the proprietary GPT-4o-mini model. Table 3 displays prompt numbers, including the optimal one (prompt 9), across zero-shot and few-shot settings (1, 3, and 5 shots); static few-shot results are omitted from the table due to consistently lower performance.

Moreover, to draw insights about the performance of the best model under zero-shot and 2-shot settings for each field, a more detailed analysis of the evaluation results with selected fields is demonstrated in Table 4.

All experiments were conducted on a machine with 2× AMD EPYC 7713 64-Core Processors, 1 TB RAM, and two NVIDIA A100 GPUs with 80 GB of memory each. The system ran Ubuntu 22.04

[8] https://huggingface.co/sentence-transformers/clip-ViT-B-32

[9] https://www.archivportal-d.de/themenportale/wiedergutmachung

with CUDA 12.4 and PyTorch 2.5. All the codes and prompts used for these experiments are publicly available on GitHub [10].

5.1 Evaluation Metrics

The quality of the models is evaluated using normalised edit distance, exact matches, and partial matches, with t denoting allowable edit distance. These metrics enable performance comparison of different models and approaches in key-value extraction.

Normalised Edit Distance. The Levenshtein (edit) distance quantifies the minimum number of single-character edits—insertions, deletions, or substitutions—required to transform one string into another [21]. To account for string length bias, we use the normalised edit distance [47], which divides the edit distance by the length of the longer string, yielding a score between 0 (identical) and 1 (completely different) [5]. In this study, we calculate the normalised edit distance (NED) between predicted and ground truth values for each key, and report the average per key and model. This enables unbiased comparison of extraction accuracy and model robustness across different keys.

Exact/Partial Match Accuracy. In KIE tasks, accuracy is assessed by how closely the predictions match the ground truth labels. Exact matches (EM) require perfect agreement in content and span, while partial matches (PM) allow for a distance of t from the correct value. This two-level evaluation captures both OCR errors and value assignment issues, providing a more nuanced assessment of model performance.

Model	Size	NED	EM (t=0)	PM (t=1)	PM (t=3)
Donut-base	-	0.415	56%	57%	58%
Donut-base-finetuned	-	0.358	59%	61%	64%
GPT-4o-mini	-	0.184	72%	76%	79%
InternVL2.0	8B	0.382	53%	55%	60%
InternVL2.0	26B	0.431	48%	52%	57%
InternVL2.0	40B	0.158	76%	79%	83%
InternVL2.0-Llama3	76B	0.286	64%	67%	70%
InternVL2.0-finetuned	40B	0.173	74%	77%	81%
InternVL2.5	8B	0.340	57%	60%	64%
InternVL2.5	26B	0.311	60%	64%	69%
InternVL2.5	38B	0.080	83%	88%	91%
InternVL2.5	78B	0.139	77%	82%	84%
InternVL2.5-finetuned	38B	0.117	79%	84%	86%

Table 2: Performance comparison of different transformer-based models on the BZKOpen dataset, including both pre-trained and fine-tuned variants. For fine-tuned models, the BZKOpen train set was used.

5.2 Scalability

As the digitisation of historical archives often involves processing millions of document images, scalability is a critical consideration for any real-world deployment of KIE systems. While MLLMs provide state-of-the-art performance, their computational cost during inference can become a significant bottleneck when applied at scale. This is particularly relevant in archival contexts, where efficient processing is necessary to meet the volume demands of large collections. To address these limitations, we evaluated alternative deployment strategies aimed at reducing inference latency without compromising model performance. The following frameworks are tested: Huggingface transformers library [44], vllm [20], and LMDeploy [6]. LMDeploy stood out as a highly effective solution, which only needs 15 minutes instead of 325 minutes (a 20x improvement in runtime) used by the transformers library to process 334 images. This acceleration is made possible through efficient GPU utilisation and parallelisation across multiple devices, making it a strong candidate for scalable deployment. The ability to parallelise processing tasks with minimal coordination overhead is essential for unlocking the full potential of MLLMs in historical document digitalisation workflows. Our findings demonstrate that with the right deployment infrastructure, even bigger MLLMs such as InternVL2.5-38B can be integrated into deployment environments requiring high-speed, high-volume document analysis.

Model	Size	Prompting Strategy	NED	EM (t=0)	PM (t=1)	PM (t=3)
InternVL2.5	38B	ZS-0	0.12	77%	83%	86%
InternVL2.5	38B	ZS-4	0.103	79%	84%	88%
InternVL2.5	38B	ZS-9	0.084	83%	88%	91%
InternVL2.5	38B	1FS-0	0.067	85%	89%	92%
InternVL2.5	38B	1FS-4	0.070	84%	89%	92%
InternVL2.5	38B	1FS-9	0.071	84%	89%	92%
InternVL2.5	38B	ZS-9	0.080	83%	88%	91%
InternVL2.5	38B	1FS-9	0.070	84%	89%	92%
InternVL2.5	38B	2FS-9	**0.060**	**86%**	**90%**	**93%**
InternVL2.5	38B	5FS-9	0.062	86%	90%	92%
GPT-4o-mini	-	ZS-9	0.184	72%	76%	79%
GPT-4o-mini	-	1FS-9	0.093	83%	87%	90%
GPT-4o-mini	-	2FS-9	0.080	84%	88%	91%
GPT-4o-mini	-	5FS-9	0.074	85%	89%	92%

Table 3: Performance comparison of zero-shot (ZS) and few-shot (FS) prompting with different numbers of shots and different prompts for the InternVL2.5-38B and GPT-4o-mini models.

Model	InternVL2.5			InternVL2.5		
Size	38B			38B		
Prompting Strategy	ZS			2FS		
Metric	EM (t=0)	PM (t=1)	PM (t=3)	EM (t=0)	PM (t=1)	PM (t=3)
BZK number	66%	73%	80%	**80%**	**89%**	**94%**
Compensation Office	41%	41%	46%	**78%**	**78%**	**80%**
Applicant First Name	**85%**	**92%**	**94%**	80%	88%	89%
Applicant Last Name	**85%**	**93%**	**94%**	80%	92%	92%
Applicant Birthdate	**91%**	**96%**	**97%**	91%	93%	96%

Table 4: Detailed per-field evaluation results for InternVL2.5-38B using zero-shot and few-shot prompting (prompt 9).

[10] https://github.com/mahsavafaie/BZK-InformationExtraction

6 Discussion

The experimental results from the preceding section offer insights into model selection, optimisation, and deployment strategies for our specific use case, generalisable to real-world KIE tasks involving historical, born-analogue sermi-structured documents.

As illustrated in Table 2, the fine-tuned Donut document transformer model achieves improved performance over its pre-trained variant; however, its results remain substantially lower than those of the MLLM models. Thus, we can say that fine-tuning alone on a rather small dataset cannot close the performance gap between document transformers and MLLMs that utilise advanced prompting strategies. On the other hand, a comparison of Table 2 and Table 1 shows that the rule-based approach outperforms the document transformer model and smaller MLLMs (8B and 26B) in accuracy. Since rule-based methods allow precise customisation and require less hardware, they are preferable for KIE tasks when document layouts are consistent and extensive rule creation is not needed.

Another key observation from Table 2 is that larger models do not necessarily provide better performance across all tasks. Our evaluation of the InternVL models reveals that both InternVL2-40B and InternVL2.5-38B outperform their larger 76B and 78B counterparts from the same series on our task. This discrepancy may be attributed to differences in the LM component among the various scales of InternVL2 and InternvVL2.5, or alternatively to the higher number of training tokens in the first stage of training, which entails training the MLP projector while keeping the vision encoder and the language model frozen [2]. These findings suggest that the quality and configuration of specific model components may have a greater impact on domain-specific KIE tasks than the total number of parameters. Notably, both InternVL2-40B and InternVL2.5-38B also surpass the performance of the commercial GPT-4o-mini model, reflecting trends observed in the OpenVLM leaderboard.

Furthermore, as shown in Table 3, iterative prompt tuning and the use of linguistic hints yield only limited impact in the few-shot setting but lead to measurable gains in the zero-shot configuration, with each iteration progressively enhancing model performance (This behaviour is also observed in the case of GPT-4o-mini). Table 3 also demonstrates that GPT-4o-mini in the few-shot setting with the optimal prompt, achieves a performance comparable to that of InternVL2.5-38B in the zero-shot setting with the best prompt. These results suggest that providing additional contextual information in the prompt, even in the absence of explicit examples, can guide the model toward improved outcomes, mirroring some of the benefits of few-shot learning, but achieved through contextual refinement rather than generalisation from observed examples. These results demonstrate that presenting the model with a single similar example for each test sample using the most basic prompt produces better results than the zero-shot strategy with the most comprehensive and informed prompt.

Finally, a fine-grained analysis of KIE performance across the keys demonstrates that for structured fields with predictable patterns—such as reference numbers or compensation authorities—few-shot prompting improves model performance by providing explicit guidance. On the contrary, for open-ended fields like names and geographical locations, few-shot examples may introduce unintended biases, making zero-shot prompting more effective. These results indicate that employing a hybrid strategy, which combines different prompting approaches depending on the expected values, can yield improved overall accuracy for such use cases.

A closer look into the failure cases did not take us to stamps and handwritings with less frequently seen fonts, but to challenging cases such as image #13 from the test set where some values are crossed out and replaced by other values, or image #29 from the test set where the card refers to multiple persecutees, but only one of them is extracted by the model.

From a deployment perspective, we find that the input image size has no impact on processing time, which simplifies scalability considerations. On the contrary, the LMDeploy implementation demonstrates a remarkable improvement in inference efficiency, achieving up to a 20x speedup over baseline setups without compromising output quality. This positions LMDeploy as a highly promising option for practical and resource-conscious deployment in MLLM-based document digitalisation pipelines.

7 Conclusion and Future Work

Currently, the 1.9 million index cards are being processed on our infrastructure with the same properties as in our experimental setting outlined in Section 5. Processing of 2,000 cards needs about one hour under the LMDeploy framework, which means the processing of 1.9 million cards needs about 40 days with InternVL2.5-38B in the zero-shot setting and with the optimal prompt. The extracted data will then be used to generate a knowledge graph. Knowledge graphs are foundational to information systems, as they enable the inference of new insights and facilitate exploratory search and discovery. Improving access to archival data increases public engagement and understanding [28]. The knowledge graph derived from the BZK cards will similarly broaden access to these historical documents, helping to uncover relationships and trends that enhance our understanding of the societal consequences of Germany's totalitarian period and the pursuit of historical justice. It is important to note that about 70% of this data will be made publicly available, as the remaining 30% falls under restricted privacy and cannot be shared for public use at the moment. The final phase involves disambiguating extracted entities and linking them to external sources for stronger querying and federated search possibilities.

Building on the findings of this study, several avenues remain open for future exploration. One promising direction involves incorporating document classification mechanisms to enable layout-specific prompting strategies. Additionally, expanding beyond German index cards to include datasets in other languages and from different domains—such as handwritten census records, legal archives, or multilingual manuscripts—would allow for a more comprehensive evaluation of model generalisability and robustness. Future experiments will also investigate the performance of emerging MLLMs to assess their scalability and effectiveness in complex, real-world archival scenarios.

Acknowledgments

This work is funded by the German Federal Ministry of Finance (*Bundesministerium der Finanzen*).

GenAI Usage Disclosure

In accordance with the ACM Policy on the use of Generative AI, we disclose that GitHub Copilot was used to assist in writing and completing portions of the source code developed during this research.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[5] William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, volume 3, pages 73–78, 2003.

[6] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023.

[7] Yihao Ding, Jean Lee, and Soyeon Caren Han. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*, 2024.

[8] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

[9] Wendy M Duff and Jessica Haskell. New uses for old records: a rhizomatic approach to archival access. *The American Archivist*, 78(1):38–58, 2015.

[10] Wendy M Duff, Andrew Flinn, Karen Emily Suurtamm, and David A Wallace. Social justice impact of archives: a preliminary investigation. *Archival Science*, 13:317–348, 2013.

[11] Robert Gaizauskas and Yorick Wilks. Information extraction: Beyond document retrieval. *Journal of documentation*, 54(1):70–105, 1998.

[12] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021.

[13] Matías García-Constantino, Katie Atkinson, Danushka Bollegala, Karl Chapman, Frans Coenen, Claire Roberts, and Katy Robson. Cliel: context-based information extraction from commercial law documents. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 79–87, 2017.

[14] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091, 2022.

[15] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.

[16] Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. Cost-effective end-to-end information extraction for semi-structured document images. *arXiv preprint arXiv:2104.08041*, 2021.

[17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[18] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without OCR. *CoRR*, abs/2111.15664, 2021. URL https://arxiv.org/abs/2111.15664.

[19] Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40, 2016.

[20] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[21] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[22] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[23] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021.

[24] Qi Li, Haijun Zhai, Louise Deleger, Todd Lingren, Megan Kaiser, Laura Stoutenborough, and Imre Solti. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *Journal of the American Medical Informatics Association*, 20(5):915–921, 2013.

[25] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1912–1920, 2021.

[26] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

[27] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. Survey of post-ocr processing approaches. *ACM Computing Surveys (CSUR)*, 54 (6):1–37, 2021.

[28] Johan Oomen, MGJ van Erp, and L Baltussen. Sharing cultural heritage the linked open data way: why you should sign up. In *Museums and the Web 2012*. 2012.

[29] OpenAI. Gpt-4o mini. https://platform.openai.com/docs/models/gpt-4o-mini, 2024. Accessed May 2025.

[30] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[31] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, et al. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*, 2023.

[32] Ricardo L Punzalan and Michelle Caswell. Critical directions for archival approaches to social justice. *The Library Quarterly*, 86(1):25–42, 2016.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[34] Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. Icdar 2019 competition on post-ocr text correction. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1588–1593. IEEE, 2019.

[35] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[36] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021.

[37] Mahsa Vafaie, Bruns Oleksandra, Nastasja Pilz, Danilo Dessí, Harald Sack, et al. Modelling archival hierarchies in practice: Key aspects and lessons learned. In *CEUR Workshop Proceedings*, volume 2981. CEUR-WS, 2021.

[38] Mahsa Vafaie, Oleksandra Bruns, Nastasja Pilz, Jörg Waitelonis, and Harald Sack. Courtdocs ontology: towards a data model for representation of historical court proceedings. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 175–179, 2023.

[39] Mahsa Vafaie, Jörg Waitelonis, and Harald Sack. Improvements in handwritten and printed text separation in historical archival documents. In *Archiving Conference*, volume 20, pages 36–41. Society for Imaging Science and Technology, 2023.

[40] Mahsa Vafaie, Mary Ann Tan, and Harald Sack. Digitalisation workflows in the age of transformer models: A case study in digital cultural heritage. 2024.

[41] Jörg Waitelonis and Harald Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, 59:645–672, 2012.

[42] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[45] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021.

[46] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[47] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007. doi: 10.1109/TPAMI.2007.1078.