

From Pixels to Structure: Lightweight Vision-Language Models for Document OCR and Structured JSON Extraction

Uddipan Basu Bir¹, Vincent Christlein¹, Andreas Maier¹, and Mathias Zinnen¹

Friedrich-Alexander-Universität Erlangen–Nürnberg, Germany
`{uddipan.bir,basu,vincent.christlein,andreas.maier,mathias.zinnen}@fau.de`

Abstract. While massive, closed-source Vision-Language Models (VLMs) set strong benchmarks for document understanding, their dependence on commercial Application Programming Interfaces (APIs) limits adoption in institutional archives due to data autonomy concerns, recurring costs, and the environmental footprint of hyperscale computing. This is especially acute in heritage digitization, where documents include historical handwriting, domain-specific terminology (e.g., jewelry, prehistory, architecture), and non-standard layouts requiring high-dimensional structured extraction. We present a comparative study of eight open-source lightweight VLMs ($\leq 7\text{B}$ parameters) for Optical Character Recognition (OCR)-to-structure across three university heritage collections. Given a document image, models must extract text and generate schema-compliant JavaScript Object Notation (JSON), enabling automatic validation and downstream use. We evaluate models under a constraint-aware protocol across zero-shot, few-shot, and fine-tuning settings, measuring extraction fidelity and structured-output quality using Character Error Rate (CER), Approximate Normalized Levenshtein Similarity (ANLS)*, and mean Average Precision F1 (mAP-F1). To assess practical improvements, we run controlled additional studies on a baseline fine-tuning pipeline, testing the independent impact of (i) hyperparameter optimization, (ii) classical image preprocessing (illumination flattening, denoising, and Contrast Limited Adaptive Histogram Equalization (CLAHE)), and (iii) multi-stage training, each compared directly to the finetune baseline. Finally, we analyze the trade-off between dataset-specific fine-tuning and a single multi-dataset checkpoint, where joint training enables one model to operate across collections but can shift performance between datasets. Overall, we show that carefully adapted $\leq 7\text{B}$ VLMs can provide a sustainable, private, high-performing alternative to manual transcription or commercial black-box systems, and we offer actionable guidance for heritage institutions seeking institution-controlled OCR-to-JSON extraction. Code for this paper is available at our project repository.¹

Keywords: Document OCR · VLMs · structured extraction · JSON generation · lightweight models · fine-tuning

¹ <https://github.com/uddipan77/Analysis-of-Lightweight-Vision-Language-Models-for-Document-OCR-and-Structured-Output-Generation>

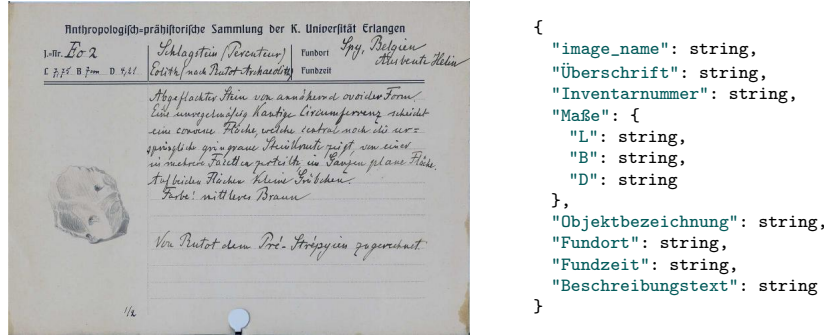


Fig. 1. Document example for the Erlangen-Prehistoric dataset (left) and its target JSON structure (right). Image from the “Inventarbuch der Anthropologisch Prähistorischen Sammlung,” courtesy of the Sammlung für Ur- und Frühgeschichte, FAU Erlangen-Nürnberg. Reproduced with permission.

1 Introduction

OCR is no longer only about reading plain text. In many real applications, the goal is to extract information from documents and return it in a structured form that can be searched, validated, or directly used by downstream systems. While standard benchmarks often focus on commercial forms like invoices or ID cards, practitioners in archives, collections, and museums face the distinct challenge of digitizing specialized historical records. In this work, we focus on OCR-to-structure: given a document image, the model should output the extracted content as a structured JSON object (see fig. 1 for a representative example from an archival inventory).

Recently, VLMs have shown strong capabilities on document understanding tasks. However, many state-of-the-art models are large, closed-source, and accessible only through commercial cloud APIs. For public archives and collections, this reliance on external providers presents significant barriers regarding data autonomy, i.e., the ability of an institution to maintain sovereignty over its own digital assets. Furthermore, the massive energy consumption of hyperscale models raises concerns about the environmental sustainability of large-scale digitization projects. In practical settings such as on-device systems, cost-sensitive cloud services, or scenarios with limited Graphics Processing Unit (GPU) resources, smaller models are often preferred. This motivates our study of lightweight VLMs: models with up to 7 billion parameters that can be deployed locally on institutional hardware. While these models are attractive for deployment, it is still unclear how reliable they are at producing correct and well-structured OCR outputs, especially when the output must follow a highly specialized archival schema.

In this paper, we present a comparative study of eight lightweight VLMs for structured OCR, evaluated on records from three specialized university collections: the Erlangen-Prehistoric dataset (prehistoric artifacts at FAU Erlangen-

Nürnberg), the Pforzheim Jewelry dataset (jewelry records at HS Pforzheim), and the Regensburg Scalalogy dataset (architectural staircase forms at OTH Regensburg). We evaluate all models in zero-shot, few-shot, and fine-tuning settings. Importantly, we follow a fair protocol that respects each model’s interface: some models only accept a fixed OCR-style instruction (e.g., a single keyword prompt such as “OCR”), while others allow custom prompts that explicitly request structured JSON extraction. Based on the overall results across zero-shot, few-shot, and fine-tuning evaluations, we select the top three models and investigate further improvements via additional studies. Finally, we analyze single-dataset versus multi-dataset fine-tuning to understand cross-collection transfer and deployment trade-offs.

We investigate practical questions that matter for institutional deployments. First, we study whether multi-stage training provides consistent gains over standard single-stage fine-tuning for OCR-to-JSON extraction. Second, we test whether hyperparameter optimization improves structured OCR performance. Third, we analyze whether classical image preprocessing can help models produce cleaner and more accurate JSON.

Overall, our goal is to provide clear guidance for practitioners from Galleries, Libraries, Archives, and Museums (GLAM) who want to use small VLMs for document OCR and structured extraction. We benchmark the models, identify strong lightweight choices, and study practical adaptation strategies that improve reliability without requiring external commercial APIs or excessive resources.

The main contributions of this paper are:

- We benchmark eight lightweight VLMs ($\leq 7\text{B}$) on three specialized archival collections, evaluating structured OCR-to-JSON as a private, sustainable alternative to commercial APIs.
- We propose a constraint-aware evaluation protocol that accounts for diverse model interfaces (fixed vs. custom prompting) in zero-shot settings.
- We evaluate the impact of fine-tuning, multi-stage fine-tuning, hyperparameter optimization, and classical image preprocessing on the top-performing models to quantify their effect on extraction accuracy and schema adherence.

2 Related Work

Traditional OCR and Document Understanding. Classical OCR systems such as Tesseract [27] rely on multi-stage pipelines (e.g., binarization, layout analysis, segmentation, recognition) and are brittle under complex layouts and degraded images. Structured extraction is often addressed by combining OCR with downstream Key Information Extraction (KIE) models. LayoutLM [32] and successors LayoutLMv2 [33] and LayoutLMv3 [15] improved multimodal document understanding, but still depend on external OCR and typically require task-specific fine-tuning with labeled bounding boxes.

End-to-End Document Understanding without OCR. To remove OCR dependence, Donut [17] maps document images directly to structured sequences, and Pix2Struct [18] pretrains on webpage screenshots to transfer to

document tasks. Nougat [7] targets academic documents (PDF-to-Markdown), while DAN [9] addresses handwriting recognition without explicit line detection. GOT-OCR 2.0 [29] unifies diverse OCR tasks and supports structured outputs (e.g., Markdown, L^AT_EX). However, these OCR-free approaches are typically evaluated primarily for accuracy rather than schema compliance required by institutional databases.

Vision-Language Models for Document Tasks. General-purpose vision-language models VLMs have also been applied to document understanding. LLaVA [21] popularized visual instruction tuning, and InstructBLIP [11] and BLIP-2 [19] improved vision-language alignment. For documents, mPLUG-DocOwl [35] applied unified instruction tuning, and Qwen-VL [3] and successors support high-resolution inputs. Recent models such as Qwen2-VL [28] and InternVL [8] have advanced performance on DocVQA [23] and FUNSD [16], but evaluations largely emphasize question answering rather than JSON validity and schema compliance. Most recently, the importance of open weights for data autonomy has been increasingly recognized. This has driven the introduction of powerful models like Pixtral [1] and DeepSeek OCR [30], which achieve state-of-the-art document understanding and allow for open deployment.

Lightweight VLMs and Efficient Adaptation. Resource constraints motivate smaller VLMs and efficient adaptation. Low-Rank Adaptation (LoRA) [14] and Quantized Low-Rank Adaptation (QLoRA) [12] enable parameter-efficient fine-tuning on accessible hardware. Lightweight VLMs include PaliGemma [6], Phi-3-Vision [24], MiniCPM-V [34], SmolVLM [22], and Florence-2 [31]. For pure, specialized text extraction, models like PaddleOCR-VL [10] have recently gained traction. However, their rigid interfaces often do not allow for the complex conversational prompting required to recognize the complex structures inherent in collection records. Furthermore, systematic comparisons of lightweight models under controlled conditions remain scarce, particularly for producing valid structured outputs (e.g., JSON) on complex archival records.

Research Gap. Despite significant progress, several gaps remain in the literature. First, most benchmarks evaluate large-scale models, leaving uncertainty about how lightweight VLMs (≤ 7 B parameters) perform on structured OCR tasks while maintaining institutional data sovereignty. Second, evaluations typically report accuracy metrics without measuring JSON validity or schema adherence, which are critical for archival integration. Third, comparisons across models rarely account for interface constraints: some models accept only fixed OCR prompts, while others support custom instructions for structured extraction. Fourth, the interaction between classical image preprocessing techniques and modern VLMs is underexplored. It is unclear whether preprocessing still benefits end-to-end models. Finally, practical adaptation strategies such as multi-stage training and hyperparameter optimization have not been systematically studied for OCR-to-JSON extraction. Our work addresses these gaps through a controlled benchmark of eight lightweight VLMs, a fair evaluation protocol respecting model constraints, and studies on preprocessing, hyperparameter optimization, and multi-stage training.

Table 1. Dataset statistics and splits. All datasets are split into 70 % training, 15 % validation, and 15 % test sets.

Dataset	Total	Train	Val	Test	Text Type	Schema Complexity
Erlangen-Prehistoric	308	216	46	46	Handwritten	Medium (8 fields)
Pforzheim Jewelry	590	413	88	89	Mostly printed	Medium (17 fields)
Regensburg Scalalogy	164	114	25	25	Mixed	High (nested, booleans)
Total	1,062	743	159	160	—	—

3 Dataset, Task, and Evaluation

3.1 Datasets and Splits

We evaluate our pipeline on three distinct datasets sourced from German university collections in archaeology, design history, and architecture. These datasets were developed in close collaboration with the respective institutions to support their ongoing efforts in digitizing historical records. They represent a spectrum of real-world archival challenges, ranging from historical handwritten text to printed catalog cards and complex, nested architectural forms. Table 1 summarizes the statistics and splits for each collection.

Erlangen-Prehistoric Dataset. This dataset originates from the Collection for Prehistory and Protohistory at the Friedrich-Alexander Universität Erlangen-Nürnberg. We processed 308 pages of the historic “Inventarbuch der Anthropologisch Prähistorischen Sammlung”, which is publicly available through the collection’s digital archive.² Ground-truth transcriptions were generated using Gemini3 and subsequently manually verified; the transcriptions are published via Zenodo.³ The documents are predominantly handwritten with printed headers or labels. Figure 1 shows an example, along with its target data structure. Each record describes archaeological or historical artifacts with fields including inventory number, dimensions, find location, find date, and descriptive text. The handwritten nature of most entries, combined with historical script variations and occasional degradation, makes this dataset particularly challenging for OCR systems.

Pforzheim Jewelry Dataset. This dataset comprises 590 digitized inventory cards from the historical teaching aids collection at the Pforzheim University, School of Design. The records come from the KUPFER research project, which investigates the collection’s history and significance.⁴ They describe jewelry from different periods and cultural areas, acquired by the Pforzheim School of Decorative Arts between 1879 and 1970. The inventory cards were derived

² <https://www.uf.phil.fau.de/abteilungen/sammlung/inventarbuecher/>

³ <https://zenodo.org/records/20668191>

⁴ KUPFER–Die Lehrmittelsammlung der Kunstgewerbeschule Pforzheim zwischen Medienarchäologie und Innovationstransfer (2023–2027), funded by the Federal Ministry of Research, Technology and Space (BMFTR) (<https://designpf.hs-pforzheim.de/kupfer>)

Gegenstand: Brosche			
Inv.-Nr.	Sch 3382/57	Museum:	Pforzheim
Standort:	1. 63		
Material:	Silber, teilweise verguldet, Perl		
Datierung:	1901		
Masse:	H 5,5 cm B 5,4 cm		
Gewicht:	4,35 g		
erworben von:	Th. Fehrer, Pforzheim, Entwerf. Ludwig Häflich	am:	1901
Beschreibung:	Brosche in Form einer Meerjungfrau, Hals, verbleibt an einem Teil des "Schwanzes" (Halskette). Die ca. 5,5 cm in Durchmesser gemessene Brosche ist aus Silber gefertigt. Der Hals ist mit einem Perlstein verziert. Der Schwanz ist aus Silber gefertigt. Der Hals ist mit einem Perlstein verziert. Der Schwanz ist aus Silber gefertigt.		
Ausstellungen:	Festschmuckausstellung Pforzheim, Berlin u.a. 1950/51 Ausstellung "Europäische Kunst"		

```
{ "Gegenstand": string,
  "Inv.Nr": string,
  "Herkunft": string,
  "Foto Notes": string,
  "Standort": string,
  "Material": string,
  "Datierung": string,
  "Masse": string,
  "Gewicht": string,
  "erworben von": string,
  "am": string,
  "Preis": string,
  "Vers.-Wert": string,
  "Beschreibung": string,
  "Literatur": string,
  "Ausstellungen": string,
  "file_name": string }
```

Fig. 2. Representative document example for the jewelry dataset (left) and its target JSON structure (right). Image courtesy of Pforzheim Jewellery Museum and HS Pforzheim (KUPFER Projekt). Reproduced with permission.

from historical books when the objects were moved on permanent loan to the Pforzheim Jewellery Museum around 1980. The task requires extracting a dense 17-field schema, including material, weight, and insurance value. An example is provided in fig. 2.

Regensburg Scalalogy Dataset. This dataset consists of 164 survey forms from the Scalalogy Collections of the Friedrich Mielke Institute (FMIS) at OTH Regensburg.⁵ See fig. 3 for an example. These records form the methodological backbone of "Scalalogy", the scientific study of stairs, and were created by Friedrich Mielke to systematically document stairs using a standardized nomenclature. The forms contain a mixture of printed field labels and handwritten entries for various staircase types (e.g., spiral and interior). With a target JSON structure featuring deeply nested objects and numerous boolean fields for technical attributes, this dataset serves as a high-complexity stress test for hierarchical schema compliance.

3.2 Task Definition and Metrics

Task Definition. Given a document image, the model must output a JSON object that follows a fixed, dataset-specific schema. Erlangen-Prehistoric dataset and Pforzheim Jewelry dataset use flat key-value dictionaries with string values, while Regensburg Scalalogy dataset uses a larger schema with nested objects and boolean fields. If a field is missing or illegible, we treat it as empty: text fields should be an empty string and boolean fields should be false. The JSON output can be deterministically rendered to Markdown for inspection and manual auditing.

Evaluation metrics. We report CER, ANLS*, and mAP-F1, and additionally track JSON validity (parsable dict outputs). Collectively, these metrics

⁵ <https://architektur.oth-regensburg.de/forschung/scalalogie>

Stadtpfarrkirche, Westfront Nordtreppe zur Empore

Haus des Hauses: Stadtpfarrkirche, Westfront Nordtreppe zur Empore
 Adresse: 1442-1513
 Bauherr:
 Baumeister:
 Bauzeit:
 Lage in/am Hause:
 Lage: außen ☐ innen ☒
 MÄßE: außen ☐ innen ☒
 MATERIAL: Holz ☒
 LAUF: rechts drehend ☒ links drehend ☐
 HANDLAUF: an der Wand / Spindel, Profil:
 STUFEN: ☒ auf steigender Tonnengewölbung ☐ auf Radialbögen ☐ beidseitig ☐ einseitig ☐
 Stütz. auf 360° ☒
 Steigung: 19 cm
 Auftritt:
 Untertritt:
 Stufenbreite:
 Stufenhöhe:
 SPINDEL: ☒ zylindrisch ☐ gewunden ☐
 p 48 cm
 ÜBERDECKUNG: Flachdecke / Wölbung
 LITERATUR:
 ABFERTIGUNG:
 11. II. 1978
 Stadt Dr. Müller

WENDETREPPEN

ORT: Bad Königshofen im Grabfeld

```
{
  "stair_type": "string",
  "Name des Hauses": "string",
  "Adresse": "string",
  "Bauherr": "string",
  "Baumeister": "string",
  "Bauzeit": "string",
  "LAGE": {
    "am Hause": bool,
    "im Hause": bool
  },
  "Gesamt Durchmesser cm": int,
  "GEHÄUSE": {
    "außen": {
      "1": bool, "2": bool, "3": bool, "4": bool, "5": bool, "6": bool
    },
    "innen": {
      "1": bool, "2": bool, "3": bool, "4": bool, "5": bool, "6": bool
    }
  },
  "MATERIAL": {
    "Holz": bool, "Ziegel": bool, "Werkstein": bool
  },
  "LAUF": {
    "rechts drehend": bool, "links drehend": bool
  },
  "Breite cm": integer,
  "Höhe cm": integer,
  "Ceiling Shape": {
    "1": bool, "2": bool, "3": bool, "4": bool
  },
  "HANDLAUF": {
    "an der Wand": bool, "an der Spindel": bool, "Profil": "string"
  },
  "STUFEN": {
    "Stück auf 360": integer,
    "Stück insgesamt": integer,
    "auf steigender Tonnengewölbung": bool,
    "auf Radialbögen": bool,
    "beidseitig": bool, "innen": bool, "außen": bool
  },
  "Notes": "string",
  "Steigung cm": integer,
  "Schweifung": "string",
  "cm Stich": integer,
  "Auftritt": {
    "cm min": integer, "cm max": integer
  },
  "Untertritt cm": integer,
  "Distanzkerbe": {
    "ja": bool, "nein": bool
  },
  "Stufenstirn": {
    "1": bool, "2": bool, "3": bool, "4": bool
  },
  "Querschnitt obere": {
    "1": bool, "2": bool, "3": bool
  },
  "Querschnitt untere": {
    "1": bool, "2": bool, "3": bool, "4": bool, "5": bool
  },
  "SPINDEL": {
    "zylindrisch": bool, "geschraubt": bool, "gewunden": bool,
    "Durchmesser des Auges cm": integer, "Durchmesser cm": integer
  },
  "Kapitell": "string",
  "BASIS": "string",
  "ÜBERDECKUNG des Gehäuses": {
    "Flachdecke": bool, "Wölbung": bool
  },
  "DEKORATION": {
    "ja": bool, "nein": bool
  },
  "Ort": "string",
  "Datum": "string",
  "Approved by": "string",
  "image_name": "string"
}
```

Fig. 3. Representative document example for the Staircase dataset (left) and its extensive target JSON structure (right). Image courtesy of the Friedrich-Mielke-Institut für Scalalogie, OTH Regensburg, Dossier “Stadtpfarrkirche Bad Königshofen im Grabfeld”. Reproduced with permission.

disentangle content quality from structured output reliability: CER and ANLS* measure how well the extracted values match the reference (even when formatting degrades), whereas mAP-F1 and JSON validity quantify whether the output is a usable JSON object under a strict parsing requirement.

Canonical ground truth and parsing policy. For each sample, the ground truth is stored as a fixed canonical JSON string (consistent formatting and a fixed key order). For a model output, we first attempt to parse the raw generation as JSON. If parsing succeeds, metrics are computed on the parsed object, and for string-based metrics we serialize the parsed prediction into the same canonical JSON format as the ground truth. If parsing fails, we compare the raw prediction text directly against the canonical ground-truth JSON string, reflecting downstream usability requirements.

CER. We compute CER against a fixed canonical ground-truth JSON string (with a consistent key order and formatting as stored). If the model output parses as JSON, we serialize the parsed prediction into the same canonical format and compute CER between the two JSON strings. If parsing fails, we compute CER between the raw model output text and the canonical ground-truth JSON string. Lower CER indicates higher overall output fidelity.

ANLS*. If both prediction and ground truth parse as JSON, we apply structured ANLS* [26] directly to the nested objects. These are then scored with ANLS similarity using threshold $\tau = 0.5$ (scores below τ are set to 0). If the prediction is not valid JSON, we fall back to a string-level ANLS-style similarity

between the raw output text and the canonical ground-truth JSON string (fixed order/format), penalizing invalid structured outputs while still giving a graded content score.

mAP-F1. Inspired by COCO-style threshold averaging [20] and DAN’s [9] mAPCER, mAP-F1 is our structure-first metric. When the prediction is valid JSON, we flatten prediction and ground truth into dot-notated leaf keys and compute per-field overlap $1 - \text{CER}(gt, pred)$ at thresholds $\theta \in \{0.3, 0.5, 0.7, 0.8\}$. We choose multiple thresholds to reflect different tolerance levels to OCR noise, so the score rewards both approximate and near-exact field matches rather than depending on a single cutoff. Empty GT/non-empty predictions count as FP; for non-empty GT, overlap $\geq \theta$ is TP, otherwise FN (and additionally FP if the prediction is non-empty). We compute precision and recall per field, macro-average them across fields to obtain $P(\theta)$ and $R(\theta)$, set $F1(\theta) = \frac{2P(\theta)R(\theta)}{P(\theta)+R(\theta)}$, and report mAP-F1 as the mean of $F1(\theta)$ over the four thresholds. If the output is not valid JSON, all non-empty GT fields are counted as FN, yielding mAP-F1 = 0.

JSON validity and parsing policy. We additionally track the JSON-validity rate, i.e., the fraction of outputs that parse (via `json.loads`) into a Python dict. For space constraints, we do not include JSON-validity columns in the main result tables. Across our experiments, JSON validity stays in the range of $\approx 95\%$ – 100% for Gemma-3, Phi-3.5-Vision, and Qwen2.5-VL across settings, but remains much lower for Donut, Florence-2, and PaddleOCR-VL even after fine-tuning. All metrics share the same initial parsing step, however, only CER and ANLS* fall back to raw-text comparison when JSON parsing fails, while mAP-F1 requires valid JSON and is set to 0 otherwise.

4 Experimental Setup and Methods

4.1 Models

We evaluate eight vision-language models (200M–7B parameters) spanning encoder-decoder and instruction-tuned architectures (Table 2). Models marked with † only accept fixed OCR-style prompts and therefore do not support few-shot demonstrations.

4.2 Constraint-aware prompting

A key challenge in benchmarking diverse VLMs is that they differ in their input interfaces and prompting flexibility. Donut-Base, Donut-CORD, Florence-2-Large, and PaddleOCR-VL are restricted to fixed OCR-style prompts or task-specific tokens (e.g., `<s_cord-v2>`, `<OCR>`) and do not support arbitrary natural-language instructions.

In contrast, Nanonets-OCR-s, Gemma-3-4B, Phi-3.5-Vision, and Qwen2.5-VL-7B accept custom prompts, enabling schema-guided structured extraction. To ensure a fair comparison, we evaluate each model under its supported interface. For

Table 2. Overview of the eight VLMs evaluated. Models marked with † only accept fixed OCR-style prompts and do not support few-shot demonstrations.

Model	Architecture Type	Parameters
Donut-Base [†] [17]	Encoder-decoder	200M
Donut-CORD [†] [17]	Encoder-decoder	200M
Florence-2-Large [†] [31]	Unified vision-language	0.77B
Nanonets-OCR-s [25]	OCR-specialized VLMs	4B
PaddleOCR-VL [†] [10]	OCR-specialized VLMs	0.9B
Gemma-3-4B [13]	Instruction-tuned VLMs	4B
Phi-3.5-Vision [24]	Instruction-tuned VLMs	4.2B
Qwen2.5-VL-7B [4]	Instruction-tuned VLMs	7B

fixed-prompt models, we apply minimal post-processing to raw generations (strip wrapper text/chat prefixes, trim whitespace and code fences) before attempting JSON parsing.

4.3 Zero-shot inference

In the zero-shot setting, we provide no in-context demonstrations. For prompt-flexible models, the input prompt includes the target JSON skeleton (i.e., the exact field structure used in the ground truth) and requests that the output follows this schema. For fixed-prompt models, we use their designated OCR prompt tokens/keywords and evaluate the resulting generations after the same minimal post-processing. All evaluations use greedy decoding (temperature = 0) for deterministic outputs, with images resized according to each model’s native preprocessing.

4.4 Few-shot inference

Few-shot evaluation is applied only to models that support custom prompting (Nanonets-OCR-s, Qwen2.5-VL, Phi-3.5-Vision, and Gemma-3). We prepend $k \in \{1, 2\}$ image-JSON demonstrations sampled from the training set:

```
[Image 1] -> {"field1": "value1", ...}
[Image 2] -> {"field1": "value1", ...}
[Query Image] ->
```

In our experiments, $k = 1$ is used for the Pforzheim Jewelry dataset, while $k = 2$ is used for the Erlangen-Prehistoric and Regensburg Scalalogy datasets.

4.5 Standard fine-tuning

We fine-tune all evaluated models. For models below 1B parameters (Donut-Base, Donut-CORD, Florence-2-Large, and PaddleOCR-VL), we perform full-parameter fine-tuning. For larger models (Nanonets-OCR-s, Gemma-3-4B, Phi-3.5-Vision, and Qwen2.5-VL-7B), we use parameter-efficient QLoRA [12] on an NVIDIA V100

(32GB Video Random-Access Memory (VRAM)), combining 4-bit quantization with LoRA adapters while keeping the base weights frozen. LoRA adapters are added to the attention and MLP layers. The best checkpoint is selected by lowest autoregressive validation CER and evaluated on the test set via autoregressive generation.

4.6 Additional Study Configurations

Using the fine-tuning pipeline described above as the baseline, we conduct three additional studies to evaluate practical strategies for improving structured OCR performance. These studies are applied only to the selected top-3 models (selection criteria reported in Section 5) and are evaluated independently to isolate each contribution.

Study 1: Hyperparameter Optimization. We compare the default fine-tuning configuration against an Optuna [2]-optimized configuration for each of the top-3 models. Optuna is used only to identify the best hyperparameters; these values are then plugged back into the same baseline fine-tuning code for the final training run. We run 20 trials using the Tree-structured Parzen Estimator (TPE) sampler, optimizing validation CER. To reduce computational cost, each trial trains for 5 epochs, after which the best hyperparameters are used for a full training run.

Study 2: Image Preprocessing. We evaluate whether classical image preprocessing improves model performance compared to using raw images. Our preprocessing pipeline consists of four sequential steps: 1. Illumination Flattening: Convert to grayscale, estimate background via morphological opening, and normalize to remove uneven lighting. 2. Denoising: Apply bilateral filtering (default) or Non-Local Means denoising to reduce noise while preserving edges. 3. CLAHE: Apply CLAHE on the luminance channel to enhance local contrast. 4. Letterbox Resize: Resize the image to a fixed canvas (e.g., 1024×1024) with padding to preserve aspect ratio. We apply the full preprocessing pipeline on top of the baseline fine-tuning setup and compare performance against the same training/evaluation run using raw images.

Study 3: Multi-stage Training. We investigate whether curriculum-based multi-stage training [5] improves convergence and final performance compared to single-stage fine-tuning. The procedure is: 1. Warm-up: Fine-tune on the target dataset’s training split for a short warm-up phase using a higher learning rate (i.e., $3 \times$ the base LR). This stage uses teacher forcing (supervised JSON targets) to quickly adapt the model to the document domain and the dataset-specific JSON style. 2. Refinement: Continue fine-tuning on the same dataset with the base learning rate for the remaining epochs. We run validation at each epoch and select the best model by the lowest validation CER (early stopping applied when validation CER stops improving for a patience period of N epochs). We compare this two-stage schedule against a single-stage baseline trained for the same total number of epochs using cosine learning-rate decay.

Table 3. Zero-shot results for all eight models across three datasets. All metrics are reported as percentages. ↓: lower is better; ↑: higher is better.

Model	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Donut-Base [†]	93.4	0.0	0.0	97.2	0.0	0.0	90.1	0.0	0.0
Donut-CORD [†]	78.8	0.0	0.0	91.9	0.0	0.0	93.1	0.0	0.0
Florence-2-Large [†]	45.6	50.4	0.0	33.6	66.4	0.0	75.6	0.0	0.0
Nanonets-OCR-s	25.7	49.4	28.5	1.9	86.5	86.5	39.3	22.8	15.2
PaddleOCR-VL [†]	69.2	20.1	0.0	51.3	57.0	0.0	78.1	0.0	0.0
Gemma-3-4B	33.6	41.8	34.6	4.4	92.1	87.4	37.6	26.3	22.5
Phi-3.5-Vision	50.9	28.1	23.4	7.3	85.3	67.9	28.7	57.1	27.9
Qwen2.5-VL-7B	21.4	54.0	56.2	2.6	96.0	96.3	21.3	67.0	33.3

Table 4. Few-shot results (1–2 examples) for models that support custom prompting and in-context demonstrations. Models restricted to fixed OCR prompts (Donut-Base, Donut-CORD, Florence-2-Large, PaddleOCR-VL) are excluded. All metrics are reported as percentages.

Model	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Nanonets-OCR-s	61.9	53.3	44.7	1.7	93.1	90.7	33.3	44.5	44.9
Gemma-3-4B	26.8	61.7	53.6	2.0	97.3	92.1	34.9	33.8	20.2
Phi-3.5-Vision	59.3	40.7	38.9	7.2	85.9	71.1	52.7	27.5	23.1
Qwen2.5-VL-7B	18.9	79.0	74.7	0.9	98.9	99.4	24.1	55.6	30.7

5 Results

5.1 Benchmark of eight VLMs Across Settings

Tables 3 to 5 summarize the performance of the evaluated models on the three datasets (Erlangen-Prehistoric, Pforzheim Jewelry, Regensburg Scalalogy) under zero-shot, few-shot (1–2 examples), and fine-tuning settings, respectively. We report CER, ANLS*, and mAP-F1. Consequently, a model can achieve low CER/high ANLS* yet still obtain a lower mAP-F1 if values are assigned to the wrong keys, fields are missing/merged, or extra fields are hallucinated. Notably, Donut can achieve lower CER/ANLS* after fine-tuning, yet still yields 0 mAP-F1 because its outputs often fail strict JSON dict parsing (schema-incompatible/malformed).

Top-3 selection. Based on the overall results across the three settings (Tables 3–5), we select the top three models, Phi-3.5-Vision, Qwen2.5-VL-7B, and Gemma-3-4B, as they show the strongest potential for reliable structured extraction (low CER, high ANLS*, and high mAP-F1) and stable JSON generation across datasets. Across zero-shot, few-shot, and fine-tuning, Phi-3.5-Vision, Qwen2.5-VL-7B, and Gemma-3-4B maintain near-saturated JSON-validity ($\approx 95\%$ – 100%) across datasets, reflecting stable structured generation. In the next subsection, we report additional studies and practical improvements on these selected models.

Table 5. Fine-tuning results for all eight models across three datasets. All metrics are reported as percentages.

Model	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Donut-Base	16.4	78.0	0.0	6.3	93.6	0.0	59.5	0.0	0.0
Donut-CORD	20.6	75.0	0.0	6.9	92.9	0.0	66.0	0.0	0.0
Florence-2-Large	10.7	75.5	45.4	11.1	89.9	0.0	57.4	4.3	0.0
Nanonets-OCR-s	23.0	82.5	81.2	0.5	96.7	87.2	20.2	79.9	50.7
PaddleOCR-VL	13.7	90.1	85.3	1.1	98.7	84.7	65.1	0.0	0.0
Gemma-3-4B	10.4	88.6	83.1	0.9	99.0	99.4	14.6	73.8	46.9
Phi-3.5-Vision	17.9	86.5	82.2	0.3	99.4	99.5	6.3	90.1	72.0
Qwen2.5-VL-7B	5.8	91.0	90.0	0.5	99.2	99.4	7.6	87.0	62.8

Table 6. Effect of hyperparameter optimization on the selected top-3 models across three datasets. Scores are reported after fine-tuning. All metrics are reported as percentages.

Model / Setting	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Gemma-3-4B (Default)	10.4	88.6	83.1	0.9	99.0	99.4	14.6	73.8	46.9
Gemma-3-4B (HPO)	9.3	88.8	65.4	0.4	99.2	99.5	3.4	92.3	77.9
Phi-3.5-Vision (Default)	17.9	86.5	82.2	0.3	99.4	99.5	6.3	90.1	72.0
Phi-3.5-Vision (HPO)	17.7	84.0	81.2	0.3	99.5	99.7	3.2	91.8	73.8
Qwen2.5-VL-7B (Default)	5.8	91.0	90.0	0.5	99.2	99.4	7.6	87.0	62.8
Qwen2.5-VL-7B (HPO)	4.3	93.4	91.2	0.5	99.3	99.5	3.1	89.4	67.1

5.2 Secondary Studies on the Selected Top-3 Models

We run three additional experiments on the selected top-3 models to test practical strategies for improving structured OCR performance. Each study is applied independently to isolate its contribution.

Study 1: Hyperparameter Optimization. Table 6 compares default fine-tuning settings against the best settings found using Optuna for each of the top-3 models.

Hyperparameter optimization generally improves performance, but the effect is not uniform across models, datasets, or metrics. The largest gains appear on the Regensburg Scalalogy dataset for all three models, while improvements on Pforzheim Jewelry dataset are comparatively small as they were already quite accurate. On Erlangen-Prehistoric dataset, hyperparameter optimization clearly benefits Qwen2.5-VL-7B, but the effect is mixed for Gemma-3-4B and Phi-3.5-Vision (e.g., some metrics improve while mAP-F1 degrade). Overall, hyperparameter optimization is a strong enhancement in several settings, but it should be evaluated empirically rather than assumed to provide consistent gains.

Study 2: Image Preprocessing. Table 7 reports the effect of classical preprocessing compared to raw images for the selected top-3 models.

Overall, image preprocessing shows a mixed, model- and dataset-dependent effect (Table 7). It improves CER for Phi-3.5-Vision on the Regensburg Scalalogy dataset (6.3% $\rightarrow$ 5.5%) and for Qwen2.5-VL-7B on the Erlangen-Prehistoric dataset (5.8% $\rightarrow$ 5.2%), but these gains do not consistently translate into better structured extraction (e.g., Phi-3.5-Vision on Regensburg Scalalogy shows lower

Table 7. Effect of image preprocessing on the selected top-3 models across three datasets. Scores are reported after fine-tuning. All metrics are reported as percentages.

Model	Preproc.	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
		CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Gemma-3-4B		10.4	88.6	83.1	0.9	99.0	99.4	14.6	73.8	46.9
Gemma-3-4B	✓	10.4	88.8	85.3	1.2	97.6	97.7	16.4	69.7	43.6
Phi-3.5-Vision		17.9	86.5	82.2	0.3	99.4	99.5	6.3	90.1	72.0
Phi-3.5-Vision	✓	19.0	80.4	76.5	0.3	99.2	99.3	5.5	89.5	61.1
Qwen2.5-VL-7B		5.8	91.0	90.0	0.5	99.2	99.4	7.6	87.0	62.8
Qwen2.5-VL-7B	✓	5.2	90.0	87.4	0.7	99.1	99.3	12.0	81.7	53.6

Table 8. Effect of multi-stage training on the selected top-3 models across three datasets. All metrics are reported as percentages.

Model	Stages	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
		CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Gemma-3-4B	single	10.4	88.6	83.1	0.9	99.0	99.4	14.6	73.8	46.9
Gemma-3-4B	multi	8.1	89.9	86.5	0.4	99.3	99.5	10.3	80.2	63.8
Phi-3.5-Vision	single	17.9	86.5	82.2	0.3	99.4	99.5	6.3	90.1	72.0
Phi-3.5-Vision	multi	18.0	86.7	83.0	0.3	99.5	99.7	4.9	90.5	72.5
Qwen2.5-VL-7B	single	5.8	91.0	90.0	0.5	99.2	99.4	7.6	87.0	62.8
Qwen2.5-VL-7B	multi	5.8	91.4	89.1	0.7	88.5	90.2	7.8	86.0	57.8

mAP-F1 after preprocessing). In several other settings preprocessing is clearly harmful—most notably on the Erlangen-Prehistoric dataset for Phi-3.5-Vision, the Pforzheim Jewelry dataset for Gemma-3-4B and Qwen2.5-VL-7B, and the Regensburg Scalalogy dataset for Gemma-3-4B and Qwen2.5-VL-7B.

Study 3: Multi-stage Training. Table 8 compares single-stage fine-tuning against two-stage curriculum training for the selected top-3 models.

In our experiments, multi-stage fine-tuning does not inherently guarantee improved accuracy; its impact is model- and dataset-dependent. From Table 8, we observe that it consistently benefits Gemma-3-4B across all three datasets. For Phi-3.5-Vision, multi-stage notably improves Regensburg Scalalogy performance, while changes on Erlangen-Prehistoric and Pforzheim Jewelry are small and can be negative depending on the metric. For Qwen2.5-VL-7B, multi-stage yields a slight gain on Erlangen-Prehistoric but degrades performance on Pforzheim Jewelry and Regensburg Scalalogy. We therefore treat multi-stage training as an empirical enhancement and report its effect comparatively rather than assuming consistent gains.

5.3 Single-Dataset vs. Multi-Dataset Fine-tuning

In addition to dataset-specific fine-tuning (one checkpoint per dataset), we fine-tune each of the selected top-3 models once on the union of all training sets (Erlangen-Prehistoric + Pforzheim Jewelry + Regensburg Scalalogy) using the same standard finetune training recipe. We then evaluate this multi-dataset checkpoint on each dataset’s test set and compare it against the corresponding single-dataset checkpoint fine-tuned only on that dataset. This comparison isolates

Table 9. Single-dataset vs. multi-dataset fine-tuning for the selected top-3 models. “Single” denotes a separate checkpoint fine-tuned per dataset; “Multi” denotes one checkpoint fine-tuned on the union of all training data and evaluated per dataset test set. All metrics are reported as percentages. ↓: lower is better; ↑: higher is better.

Model	Erlangen-Prehistoric			Pforzheim Jewelry			Regensburg Scalalogy		
	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑	CER↓	ANLS*↑	mAP-F1↑
Gemma-3-4B (Single)	10.4	88.6	83.1	0.9	99.0	99.4	14.6	73.8	46.9
Gemma-3-4B (Multi)	8.7	90.6	87.9	0.3	99.4	99.7	9.1	81.5	68.0
Phi-3.5-Vision (Single)	17.9	86.5	82.2	0.3	99.4	99.5	6.3	90.1	72.0
Phi-3.5-Vision (Multi)	19.7	84.0	45.4	0.5	99.2	99.4	7.4	84.3	54.1
Qwen2.5-VL-7B (Single)	5.8	91.0	90.0	0.5	99.2	99.4	7.6	87.0	62.8
Qwen2.5-VL-7B (Multi)	7.1	89.3	91.9	1.6	98.2	99.1	24.9	65.2	41.4

Zero-shot prediction

```
{
  "Überschrift": "Anthropologisch-prähistorische ...",
  "Inventarnummer": "J.-Nr. E024",
  "Maße": {"L": "", "B": "", "D": ""},
  "Objektbezeichnung": "Primitiver Kratzer",
  "Fundort": "Spay, Belgien",
  "Fundzeit": "Aumante Helin",
  "Beschreibungstext": "Flacher Stein ... [many OCR errors]"
}
```

Fine-tuned prediction

```
{
  "Überschrift": "Anthropologisch-prähistorische ...",
  "Inventarnummer": "Eo 24",
  "Maße": {"L": "6,25", "B": "3,95", "D": "1,7"},
  "Objektbezeichnung": "Primitiver Kratzer ...",
  "Fundort": "Spy, Belgien Ausbeute Helin",
  "Fundzeit": "",
  "Beschreibungstext": "Flacher Stein ... [partially correct]"
}
```

Fig. 4. Qualitative comparison for Gemma-3-4B on an Erlangen-Prehistoric sample. Fine-tuning corrects the inventory number and measurements and reduces sample-level CER from 26.8 % to 13.9 %, while longer handwritten text and **Fundort**/**Fundzeit** separation remain challenging.

whether multi-dataset training improves generalization or introduces negative transfer across distinct schemas and text modalities.

Overall, the effect of multi-dataset fine-tuning is clearly model-dependent rather than a uniform trade-off. For Gemma-3-4B, the multi-dataset checkpoint consistently outperforms the corresponding single-dataset checkpoints on all three datasets, improving both CER and ANLS* (Table 9). In contrast, Qwen2.5-VL-7B shows the opposite behavior: the multi-dataset checkpoint generally degrades relative to single-dataset fine-tuning, with worse CER and ANLS* on all three datasets and the strongest drop on Regensburg Scalalogy (CER 7.6 % → 24.90 %, ANLS* 87.0 % → 65.2 %). While mAP-F1 slightly improves on Erlangen-Prehistoric (90.0 % → 91.9 %), it decreases on Pforzheim Jewelry (99.4 % → 99.1 %) and Regensburg Scalalogy (62.8 % → 41.4 %). Phi-3.5-Vision also degrades under multi-dataset fine-tuning on all three datasets, but the effect is markedly smaller than for Qwen, especially on Pforzheim Jewelry (CER 0.30 % → 0.50 %, ANLS* 99.4 % → 99.2 %). We therefore do not observe a single general rule for multi-dataset fine-tuning in this benchmark; instead, it can either improve cross-dataset robustness (Gemma) or introduce negative transfer of varying severity (Phi, Qwen) depending on the model.

Qualitative example. Figure 4 illustrates a representative Erlangen-Prehistoric example for Gemma-3-4B, where fine-tuning improves field binding for inventory number and measurements, while some handwritten text and field-boundary errors remain.

6 Conclusion

We benchmarked eight lightweight VLMs for document OCR with schema-guided JSON extraction across three archival datasets covering handwritten inventory books, mostly printed jewelry records, and structurally complex staircase forms. Overall, instruction-tuned general-purpose VLMs (Qwen2.5-VL, Phi-3.5-Vision, and Gemma-3) are the strongest candidates, outperforming several OCR-specialized alternatives in structured generation. This indicates that reliable extraction depends not only on text recognition, but also on instruction following, field binding, and stable schema-conformant decoding.

Few-shot prompting is not consistently beneficial: it can help on cleaner printed records such as Pforzheim Jewelry, but remains mixed or negative on handwritten and structurally complex datasets, where longer prompts can increase instruction complexity. Fine-tuning is the most effective adaptation strategy, with full fine-tuning for sub-1B models and QLoRA for larger models enabling strong local performance on a single consumer GPU. Among the additional studies, hyperparameter optimization provides the most reliable modest gains, particularly on the more difficult Regensburg Scalalogy dataset, whereas image preprocessing and multi-stage training are strongly model- and dataset-dependent and can also degrade structured extraction.

The single- versus multi-dataset experiments further show that joint training is not uniformly beneficial: Gemma-3-4B improves across all three collections, suggesting better cross-schema robustness, whereas Phi-3.5-Vision and especially Qwen2.5-VL-7B show signs of negative transfer. This highlights that deployment decisions should consider whether an institution needs one general checkpoint across collections or separate dataset-specific checkpoints for maximum accuracy.

Our study has two main limitations: we evaluate hyperparameter optimization, preprocessing, and multi-stage training independently, and the lightweight VLM landscape evolves rapidly. Thus, our findings should be read as guidance for the evaluated model families rather than an exhaustive statement about all emerging OCR/document models. Future work should explore constrained decoding for valid JSON, layout-aware field binding, handwriting-specific adaptation, broader multilingual datasets, and controlled comparisons of full fine-tuning versus QLoRA under matched compute budgets.

Disclosure of Interests The authors declare no competing interests.

Acknowledgments This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) within the project Sammlungen, Objekte, Datenkompetenzen (SODa), which aims to establish a data competence center for university collections. The authors also gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b268dc. NHR funding is provided by federal and Bavarian state authorities. We extend our gratitude

to Tabea Schmid and Katharina Wittemann from the Faculty of Design at HS Pforzheim, Sophie Schlosser from the Friedrich-Mielke Institute for Scalology at the OTH Regensburg, as well as Doris Mischka and Thorsten Uthmeier from the Institut für Vor- und Frühgeschichte at the FAU Erlangen. Their domain expertise, constructive criticism, and support were invaluable for our experiments.

References

1. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., et al.: Pixtral 12b. arXiv preprint arXiv:2410.07073 (2024) [4](#)
2. Akiba, T., Sano, S., Yanase, T., et al.: Optuna: A next-generation hyperparameter optimization framework. In: KDD (2019) [10](#)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) [4](#)
4. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025) [9](#)
5. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009) [10](#)
6. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: PaliGemma: A versatile 3b VLM for transfer. arXiv preprint arXiv:2407.07726 (2024) [4](#)
7. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. In: ICLR (2024), poster [4](#)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024) [4](#)
9. Coquenat, D., Chatelain, C., Paquet, T.: DAN: A segmentation-free document attention network for handwritten document recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8227–8243 (2023) [4](#), [8](#)
10. Cui, C., Sun, T., Liang, S., et al.: PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528 (2025) [4](#), [9](#)
11. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023) [4](#)
12. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient fine-tuning of quantized LLMs. In: NeurIPS (2023) [4](#), [9](#)
13. Google DeepMind: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025) [9](#)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) [4](#)
15. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-training for document AI with unified text and image masking. In: ACM International Conference on Multimedia (2022) [3](#)
16. Jaume, G., Ekenel, H.K., Thiran, J.P.: FUNSD: A dataset for form understanding in noisy scanned documents. In: ICDARW (2019) [4](#)
17. Kim, G., Hong, T., Yim, M., et al.: OCR-free document understanding transformer. In: ECCV (2022) [3](#), [9](#)

18. Lee, K., Joshi, M., Turc, I., et al.: Pix2struct: Screenshot parsing as pretraining for visual language understanding. In: ICML (2023) 3
19. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) 4
20. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: Common objects in context. In: ECCV (2014) 8
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 4
22. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Ben Allal, L., Lozhkov, A., Tazi, N., Srivastav, V., Lochner, J., Larcher, H., Morlon, M., Tunstall, L., von Werra, L., Wolf, T.: SmolVLM: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025) 4
23. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A dataset for VQA on document images. In: WACV (2021) 4
24. Microsoft: Phi-3.5-vision-instruct (model card) (2024), <https://huggingface.co/microsoft/Phi-3.5-vision-instruct>, accessed 2026-02-23 4, 9
25. Nanonets: Nanonets-OCR-s: A lightweight OCR model for structured document extraction (2024) 9
26. Peer, D., Schöpf, P., Nebendahl, V., Rietzler, A., Stabinger, S.: ANLS* – a universal document processing metric for generative large language models. arXiv preprint arXiv:2402.03848 (2024) 7
27. Smith, R.: An overview of the Tesseract OCR engine. In: ICDAR 2007. IEEE (2007) 3
28. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 4
29. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., Han, C., Zhang, X.: General OCR theory: Towards OCR-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024) 4
30. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025) 4
31. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: CVPR (2024) 4, 9
32. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of text and layout for document image understanding. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020) 3
33. Xu, Y., Xu, Y., Lv, T., et al.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: ACL (2021) 3
34. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhao, H., Li, Y., He, H., Liu, Z., Feng, Y., et al.: MiniCPM-V: A GPT-4V level MLLM on your phone. arXiv preprint arXiv:2408.01800 (2024) 4
35. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al.: mPLUG-DocOwl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499 (2023) 4