Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Pixels to Pairs: A Comprehensive Benchmark of LLM-Driven Key–Value Extraction in Noisy Document Settings

Zahra Anvari

zahra.anvari.uta@gmail.com

Westcliff University

Research Article

Keywords: Document AI, Large Language Models, Key–Value Extraction, Optical Character Recognition, In-Context Learning, Document Understanding

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11199953/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Pixels to Pairs: A Comprehensive Benchmark of LLM-Driven Key–Value Extraction in Noisy Document Settings

Zahra Anvari[1]

[1]College of Technology & Engineering, Westcliff University, San Francisco, CA, USA.

Contributing authors: zahraanvari@westcliff.edu;

Abstract

Large language models (LLMs) have demonstrated strong capabilities in document key–value pair (KVP) extraction, yet controlled evaluations of their robustness to optical character recognition (OCR) output remain limited. This leaves an important gap in understanding their reliability within real-world OCR-to-LLM pipelines. Unlike end-to-end Vision-Language Models (VLMs) that conflate visual perception with semantic extraction, modular pipelines allow errors to be explicitly isolated and audited. To bridge this evaluation gap, we introduce a controlled, large-scale benchmarking framework designed to distinguish downstream LLM extraction behavior from upstream OCR degradation. This benchmark evaluates 136 experimental configurations and 17,688 deterministic document-level inference instances across five instruction-tuned open-weight LLMs (2B–8B parameters), three complementary datasets, and four text-quality conditions. This extensive matrix pairs a full zero-shot comparison and targeted one- to three-shot experiments with a confirmatory sensitivity analysis comprising 40 configurations across 20 frozen demonstration sets. By separating Key Recall (annotated-field recovery) from Exact Match and Value F1 (exact and partial value recovery, respectively), we reveal whether OCR degradation affects field identification and value reproduction differently across models. Crucially, our findings challenge three practical evaluation assumptions: (1) that clean-text performance is a reliable proxy for real-world robustness, (2) that model rankings remain consistent across annotation-derived Gold and OCR-derived text conditions, and (3) that providing few-shot demonstrations monotonically improves extraction accuracy. By exposing vulnerabilities such as model-ranking reversals and unstable few-shot gains, this benchmark provides a rigorous and reproducible framework for evaluating LLM reliability under noisy document conditions. To facilitate reproducible research, our entire benchmarking framework, dataset splits, and evaluation scripts are open-sourced.

Keywords: Document AI, Large Language Models, Key–Value Extraction, Optical Character Recognition, In-Context Learning, Document Understanding

1 Introduction

Key-value pair (KVP) extraction is central to Document AI (DocAI) applications such as invoice processing, receipt understanding, and form digitization [10, 12]. Instruction-tuned large language models (LLMs) offer a flexible, text-only alternative to task-specific extraction models [1, 6, 13], framing information extraction as structured generation without task-specific fine-tuning. However, controlled evaluations of their robustness to optical character recognition (OCR)-derived input remain limited. Recent surveys indicate that evaluations of LLM-based document understanding frequently use clean, natively digital, or annotation-derived

textual inputs, leaving robustness to OCR-derived text comparatively underexplored [3].

Addressing this evaluation gap is critical for practical deployment. While recent advancements in Vision-Language Models (VLMs) such as GPT-4, LLaVA, and mPLUG-DocOwl [8, 18, 20] offer end-to-end document understanding capabilities, decoupled OCR-to-LLM pipelines remain practically relevant in enterprise systems handling sensitive data, such as financial invoices or medical records. These modular pipelines remain attractive for two distinct reasons: they permit the independent selection of OCR and extraction components, and they allow the downstream semantic reasoning to be handled by smaller instruction-tuned models (such as the 2B–8B models evaluated here) that run on a single GPU. Because end-to-end VLMs jointly perform visual recognition, OCR-like transcription, layout reasoning, and semantic extraction within a single opaque architecture, they obscure the exact origin of failures. Including them in the same comparison would therefore conflate upstream perception with downstream extraction, preventing performance changes from being attributed specifically to textual fidelity, the central experimental variable in this study. Instead, we adopt a controlled component-level design that standardizes downstream inference and evaluation while varying the input text condition.

In practice, document degradations such as blur, noise, shadows, and poor illumination severely reduce OCR reliability [2]. Recognition errors, token fragmentation, numeric corruption, and altered reading order flatten a two-dimensional document into noisy text, removing spatial cues and disrupting the LLM's ability to associate keys with values. We therefore organize the study around three evaluation assumptions:

1. **Clean-Text Robustness Proxy:** Models achieving high accuracy on clean text are proportionally robust to upstream OCR noise.
2. **Rank Stability:** A model that outperforms its peers on digital documents will maintain its superiority on noisy, scanned documents.
3. **The Few-Shot Guarantee:** Providing in-context demonstrations consistently and monotonically improves extraction accuracy in noisy environments.

To empirically investigate these assumptions, we introduce a systematic benchmarking framework designed to distinguish downstream LLM extraction behavior from upstream OCR degradation. Because LLM performance is highly sensitive to prompt configuration, OCR engine, and output parsing, drawing robust conclusions requires evaluating models across a broad permutation of conditions. We achieve this by evaluating Form Understanding in Noisy Scanned Documents (FUNSD) [12], Scanned Receipt OCR and Information Extraction (SROIE) [10], and CORD [21] datasets using annotation-derived (Gold) text and outputs from Tesseract [25], EasyOCR [11], and PaddleOCR [5]. Our large-scale experimental matrix shown in Figure 1 executes 136 configurations across five open-weight models from the Gemma [6], Mistral [13], Qwen2.5 [22], and LLaMA 3 [1] families. It combines zero-shot evaluation of all five models, targeted one- to three-shot evaluation of Qwen2.5–7B and LLaMA 3–8B, and 40 confirmatory sensitivity configurations, yielding 17,688 deterministic document-level inference instances. Our evaluation uses complementary metrics rather than a single aggregate score. **Key Recall** measures annotated-field recovery, while **Exact Match (EM) and Value F1 (VF1)** measure exact and partial value recovery. This decomposition reveals whether OCR degradation affects field identification and value reproduction differently, without interpreting the metrics as direct measures of hallucination frequency. In summary, our primary contributions challenge the preceding assumptions and provide a rigorous, reproducible framework for evaluating in-context learning under noisy document conditions:

- **A Large-Scale Evaluation Framework:** We present a comprehensive controlled benchmark comprising 136 configurations and 17,688 deterministic document-level inference instances across multiple document types, text conditions, models, and prompting settings.
- **Granular Metric Decomposition:** We separate annotated-field recovery, measured by Key Recall, from exact and partial value recovery, measured by EM and VF1, enabling granular analysis of how OCR degradation affects different aspects of extraction.
- **Exposing the Clean-Text Fallacy (Refuting Assumptions 1 & 2):** We demonstrate that clean-text performance does not uniformly predict OCR-derived performance: relative to the corresponding Gold condition, the top-ranked zero-shot model changes in 10

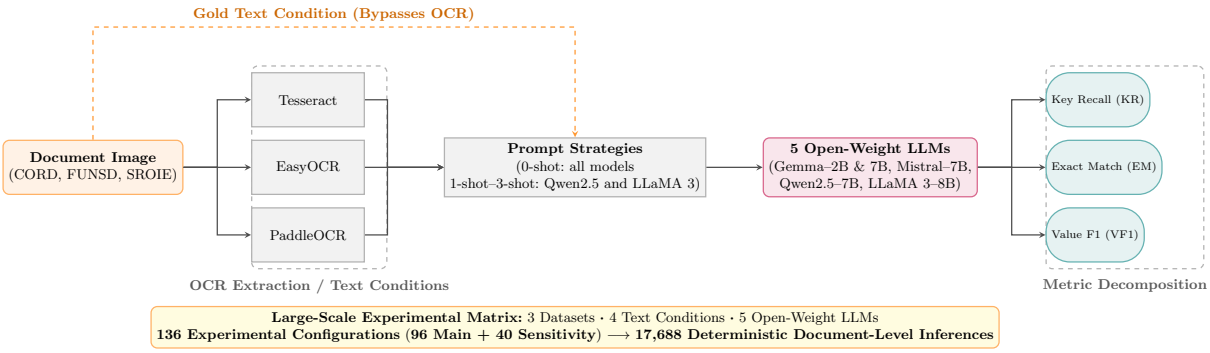


Fig. 1 The end-to-end benchmark architecture pipeline. A document image is processed through branching OCR engines (or a Gold Text bypass) to construct a prompt. The resulting text is fed into the LLMs, and the predictions are evaluated using three complementary metrics, Key Recall (KR), Exact Match (EM), and Value F1 (VF1), to distinguish annotated-field recovery from exact and partial value recovery. The bottom banner highlights the combinatorial scale of the executed experimental matrix.

of 27 dataset–OCR–metric comparisons, as detailed in Section 6.7.

- **Unstable Few-Shot Dynamics (Refuting Assumption 3):** We reveal that in-context demonstrations do not guarantee monotonic accuracy improvements under OCR noise, exhibiting high sensitivity to demonstration complexity and composition.
- **Open-Source Reproducibility:** We release our entire benchmarking suite, including all configurations, dataset splits, and evaluation tools, providing the community with a framework to evaluate future LLMs. Code is available at https://github.com/zanvari/pixels-to-pairs.

2 Related Work

2.1 Document Key-Value Pair Extraction and Multimodal Models

Early document extraction systems relied on templates, regular expressions, and spatial heuristics [7, 23]. Deep learning introduced representations that jointly model text and layout, including Chargrid [14] and the FUNSD benchmark [12]. Transformer-based models subsequently integrated textual, spatial, and visual information: LayoutLM and its successors [9, 29, 30], DocFormer [4], FormNet [16], StructuralLM [17], and UDOP [26] established strong supervised document-understanding approaches. Generative models such as Donut [15], DocVLM [24], and LayoutLLM [19] incorporate document images more directly. These methods demonstrate the value of visual and spatial information but differ from our OCR-centered text-only setting, which isolates language-model behavior after document content has been converted to text. A broader review of multimodal document extraction trends is provided in [3].

2.2 LLMs for Structured Information Extraction

Instruction-tuned LLMs can infer structured relationships from text without task-specific fine-tuning. Mistral [13], Gemma [6], Qwen2.5 [22], and LLaMA 3 [1] provide zero-shot and few-shot capabilities, while instruction tuning [28] and frameworks such as InstructUIE [27] demonstrate unified prompt-based interfaces for information extraction. Comparisons remain difficult, however, because studies vary in modality, supervision, prompting, decoding, and post-processing. The interaction between OCR transcription quality and prompt-based text-only KVP extraction is particularly under-characterized across multiple OCR engines, LLM families, and document structures.

2.3 OCR Noise and Downstream Text Degradation

Document degradations can produce character substitutions, token fragmentation, line reordering, and numeric corruption during OCR [2], with direct consequences for KVPs containing

dates, monetary values, identifiers, and addresses. We evaluate outputs from Tesseract [25], EasyOCR [11], and PaddleOCR [5] under matched annotation-derived and OCR-derived conditions. By holding downstream inference and evaluation fixed across three document benchmarks and zero- to three-shot prompting, the study isolates how upstream textual fidelity interacts with document structure, model behavior, and demonstration selection.

3 Datasets

We evaluate three document-understanding benchmarks: FUNSD forms, SROIE receipts, and CORD receipts (Fig. 2). They differ in layout dependence, annotation schema, textual density, and numeric content, enabling comparison across structurally distinct extraction settings.



(a) FUNSD (b) SROIE (c) CORD

Fig. 2 Example documents from the evaluated test splits: (a) FUNSD [12], (b) SROIE [10], and (c) CORD [21].

3.1 CORD

CORD [21] contains retail receipts with variable length, layout, and item density. Numerous line items and repeated monetary values make key–value alignment challenging, while OCR digit corruption can directly affect value recovery. CORD therefore represents a dense, numerically intensive receipt setting.

3.2 FUNSD

FUNSD [12] contains scanned forms whose key–value relationships often depend on two-dimensional spatial organization. Converting these forms to plain text removes layout cues and can obscure associations between keys and distant values, making FUNSD particularly challenging for a text-only extraction pipeline.

3.3 SROIE

SROIE [10] contains relatively short receipts annotated with four canonical fields: *company*, *date*, *address*, and *total*. Its compact schema reduces structural ambiguity, but OCR changes in addresses and numeric values can strongly affect exact and partial value recovery. Because only the annotated field inventory is scored, additional semantically meaningful receipt fields are outside the primary evaluation.

Together, FUNSD emphasizes layout-dependent associations, SROIE evaluates a compact receipt schema, and CORD stresses extraction from longer, numerically dense receipts. Dataset-specific construction of the annotation-derived text used in the benchmark is described in Section 4.

4 Methodology

This section defines the KVP extraction task and the controlled experimental protocol used across models, datasets, prompting regimes, and text conditions.

4.1 Task Formulation

Given a document image I, a text representation T is obtained either from dataset annotations or from an OCR engine. The objective is to recover

$$\mathcal{P} = \{(k_i, v_i)\}_{i=1}^{N},$$

where k_i is a semantic key (e.g., `TOTAL`, `DATE`) and v_i its associated value. OCR may introduce character substitutions, token fragmentation, numeric corruption, and reading-order errors; flattening forms into text can additionally remove spatial cues needed to associate keys and values.

4.2 Models Evaluated

The benchmark evaluates five instruction-tuned, open-weight decoder-only checkpoints without task-specific fine-tuning: Gemma-2B-it and Gemma-7B-it [6], Mistral-7B-Instruct-v0.2 [13], Qwen2.5-7B-Instruct [22], and LLaMA-3-8B-Instruct [1]. These checkpoints were selected to provide coverage across four widely used model families and the practical 2B–8B parameter range while remaining feasible for single-GPU academic experimentation.

4.3 Prompting Strategy

All models use a unified prompt that defines the extraction task, specifies the required JSON output structure, and discourages unsupported fields. The overall template is shared across datasets, with only schema-dependent wording changed; we do not perform model-specific prompt optimization.

All models are evaluated zero-shot. Qwen2.5–7B and LLaMA 3–8B are also evaluated with one, two, and three demonstrations drawn exclusively from the corresponding training split and disjoint from evaluation documents.

For FUNSD, demonstration inputs use the same spatially reconstructed annotation text as the Gold text condition (Section 4.6), and targets are generated directly from annotated question–answer links. Selected demonstration texts and targets were verified against the training annotations. The same clean training demonstrations are held fixed across the evaluated Gold and PaddleOCR few-shot conditions, so the text source varies only for the target document.

All datasets share the same JSON output contract, but the extraction scope follows the benchmark annotation structure. The SROIE prompt explicitly enumerates its four canonical fields i.e. company, address, date, and total, whereas the FUNSD and CORD prompts request all reasonable document-grounded KVPs because their evaluated field inventories are document dependent.

4.4 Output Parsing and Canonicalization

Outputs are parsed deterministically and without semantic correction. We first parse the complete response as the requested JSON object. If this fails, a conservative recovery step retains only complete, independently decodable dictionary objects with the expected structure. Incomplete objects are not repaired, and no regex extraction, document-text fallback, ground-truth-aware recovery, or semantic heuristics are used. Unparseable responses are treated as empty predictions.

Formatting-tolerant key normalization is applied only during scoring (Section 5); it does not use fuzzy matching, stemming, synonym expansion, or semantic key mapping.

4.5 Datasets and Evaluation Splits

The benchmark uses FUNSD forms, CORD receipts, and SROIE receipts. Few-shot demonstrations come only from the corresponding training data and are disjoint from evaluation documents. FUNSD contains 50 original test documents; three contained no evaluable ground-truth KVPs under the benchmark preprocessing and evaluation protocol, leaving 47 evaluable documents. The same 47-document subset is used across all FUNSD models, OCR conditions, and prompting configurations. CORD and SROIE contain 100 and 347 evaluated documents, respectively. The excluded documents from FUNSD test split were 85201976, 85540866, and 85629964; each yielded zero annotated ground-truth KVPs after extracting valid question–answer links, making KR, EM, and VF1 undefined.

4.6 Text Input Construction and Conditions

Each model is evaluated under a clean annotation-derived text condition and three OCR-derived conditions. Because the datasets provide different annotation structures, clean text is constructed using a dataset-specific but deterministic procedure.

4.6.1 FUNSD

FUNSD annotations provide text blocks and bounding boxes $[x_0, y_0, x_1, y_1]$. After removing empty blocks and invalid boxes, we group blocks into lines when their vertical overlap is at least 0.50 or their vertical-center distance is at most 0.40 times the maximum of the block height and the candidate line's median block height. Multiple qualifying lines are resolved using overlap and normalized center distance; blocks are sorted left to right within lines and lines top to bottom by median center, with minimum x_0 breaking ties. Spaces join blocks and newlines separate lines. These rules were fixed before the corrected downstream runs and were not tuned against KVP performance. We call the result *spatially reconstructed Gold text*: a deterministic approximation rather than a uniquely canonical reading order.

4.6.2 CORD

CORD clean text is reconstructed from its structured annotations by traversing the annotated

valid lines and their constituent words in the order provided by the dataset and concatenating the annotated word text. We audited the resulting clean-text files against this reconstruction procedure and obtained exact agreement for all 100 evaluated documents.

4.6.3 SROIE

SROIE clean text is constructed from the dataset's provided OCR box–transcription annotations while preserving the annotated document-text ordering. An audit against an independent spatial reconstruction showed equivalent token content for 346 of 347 documents. The sole discrepancy was an encoding/replacement-character artifact in the source annotation; the existing clean transcription was retained rather than introducing the corrupted character into the reference text.

4.6.4 OCR-Derived Text

For all engines, the LLM received the engine's native text sequence; bounding boxes were neither included nor geometrically re-sorted. Tesseract output came directly from `pytesseract.image_to_string` with English recognition. EasyOCR `readtext(..., detail=0)` strings and PaddleOCR 3.x `rec_texts` were joined with newline characters in API order; PaddleOCR orientation classification, document unwarping, and text-line orientation were disabled. Thus, the OCR inputs retain engine-specific recognition, tokenization, line-grouping, and reading-order effects and represent a native plain-text rather than geometry-aware OCR-to-LLM pipeline.

4.7 Implementation and Inference Settings

All experiments use deterministic greedy decoding (`do_sample=False`), standard model chat templates when available, and no model-specific prompt tuning, reranking, semantic post-processing, or decoding modifications. The unified Hugging Face `transformers` pipeline processes documents independently and stores raw generations, parsed predictions, per-document scores, prompt lengths, truncation flags, generated token counts, and output-limit events. Fixed token budgets, software versions, and hardware are reported in Section 4.9.

4.8 Confirmatory Sensitivity Analysis

Because the primary few-shot experiments showed non-monotonic responses to additional demonstrations, we conducted a confirmatory analysis of *demonstration composition*. It uses FUNSD spatially reconstructed Gold text with Qwen2.5–7B and LLaMA 3–8B and fixes the demonstration count at three, so only composition varies.

4.8.1 Candidate Construction

Candidates came exclusively from the FUNSD training split and required (i) nonempty corrected spatial-Gold text, (ii) at least one linked question–answer pair, and (iii) nonempty linked keys and values. This yielded 106 eligible documents. Eight documents used in earlier pilot sets were excluded to keep the confirmatory analysis independent, leaving 98 candidates; no test document was used for selection.

4.8.2 Structural Complexity Profile

Each eligible training document was characterized using eight predefined training-only structural features: (1) number of annotation nodes, (2) number of linked KVPs, (3) number of unique normalized keys, (4) number of keys occurring more than once, (5) number of repeated-key occurrences beyond the first, (6) mean value length in words, (7) maximum value length in words, and (8) number of words in the spatially reconstructed Gold text.

Each feature was converted to its percentile rank over eligible training documents, and document complexity was defined as the mean percentile rank:

$$C(d) = \frac{1}{8} \sum_{j=1}^{8} R_j(d), \tag{1}$$

where $R_j(d)$ is the percentile rank for feature j. Thus all eight dimensions contribute equally, and the score uses only training-document structure and text statistics—never model predictions or test performance.

The 25th and 75th percentiles of the resulting training complexity distribution were $Q_{25} = 0.3578$ and $Q_{75} = 0.6557$, respectively. These thresholds were computed from the eligible training-document complexity distribution, before any model inference or inspection of test performance. After the pilot-document exclusions, the candidate pool contained 26 low-complexity documents ($C(d) \leq Q_{25}$), 46 representative documents ($Q_{25} < C(d) < Q_{75}$), and 26 high-complexity documents ($C(d) \geq Q_{75}$). The terms *Low*, *Representative*, and *High* therefore denote algorithmically defined strata of this training-only complexity distribution; they are not human-assigned difficulty labels and do not encode whether an LLM succeeds or fails on a document.

4.8.3 Frozen Demonstration Sets

We constructed four demonstration-selection strategies: *Low*, containing three low-complexity examples; *Representative*, containing three examples from the middle complexity range; *High*, containing three high-complexity examples; and *Diverse*, containing one Low, one Representative, and one High example.

Five independent three-document sets were generated per strategy (20 sets total). Selection was fully algorithmic and reproducible. Within each stratum, candidate identifiers were first sorted by document identifier and then permuted using sequential draws from a single NumPy random-number generator initialized with fixed seed 20260915. The first 15 permuted documents in each stratum were partitioned consecutively into five homogeneous three-shot sets. The next five documents from each of the Low, Representative, and High strata were combined position-wise to form five Diverse sets, each containing one document from each stratum.

No document was reused across the 20 confirmatory sets, yielding 60 unique demonstrations. No human annotator assigned documents to complexity strata or selected examples based on perceived difficulty, representativeness, model behavior, or expected usefulness as demonstrations. Manual inspection was limited to a source-integrity audit verifying that the already selected documents had valid annotation and spatial-Gold sources, nonempty text, and valid linked KVPs; this audit did not alter the complexity labels or use model outputs. The complete set manifest, including document identifiers, strata, complexity definition, quartile thresholds, and random seed, was frozen before confirmatory inference. No set or document was subsequently selected, replaced, reordered, or modified on the basis of downstream KR, EM, VF1, model outputs, or test-set behavior.

4.8.4 Confirmatory Evaluation

The 20 frozen sets were evaluated independently with both models, yielding 40 configurations and 1,880 document-level inference instances over the same 47 FUNSD test documents. Prompt structure, corrected spatial-Gold input, decoding, and evaluation match the main benchmark. Results are summarized across valid sampled sets within each strategy rather than selecting the best-performing set.

4.9 Experimental Matrix and Computational Setup

For each dataset, five models are evaluated zero-shot under Gold, PaddleOCR, EasyOCR, and Tesseract inputs (20 configurations). Qwen2.5–7B and LLaMA 3–8B additionally use one-, two-, and three-shot prompting under Gold and PaddleOCR (12 configurations), giving 32 configurations per dataset and 96 overall. With 47 FUNSD, 347 SROIE, and 100 CORD documents, the main matrix comprises 15,808 document-level inference instances. Adding the 40 confirmatory FUNSD configurations and their 1,880 instances yields 136 experimental configurations and 17,688 document-level inference instances.

All final runners use an 8,192-token input budget, at most 1,024 generated tokens, batch size one, deterministic greedy decoding (`do_sample=False`), standard chat templates when available, and fixed NumPy/PyTorch seeds. Generation parameters are not tuned separately by model.

The final environment used Python 3.13.15, PyTorch 2.11.0 with CUDA 12.8, Transformers 5.16.1, and an NVIDIA A100 GPU with 40 GB memory. Models were run sequentially and document-level artifacts persisted for resumable execution. Cache cleanup between model blocks, when required, did not alter model weights, prompts, generation settings, or saved outputs.

5 Evaluation Methodology

Predictions are evaluated with the same deterministic normalization, matching, and scoring procedure under every text condition. The models receive only the supplied text; no image, bounding boxes, or layout coordinates.

5.1 Evaluation Scope

For ground truth

$$\mathcal{P} = \{(k_i, v_i)\}_{i=1}^{N}$$

and model prediction $\hat{\mathcal{P}}$, evaluation is reference-centered: Key Recall (KR), Exact Match (EM), and Value F1 (VF1) quantify recovery of the N annotated KVPs. Predictions whose normalized keys do not correspond to an annotated key do not contribute to these metrics; additional fields are considered descriptively during output inspection. Thus, the primary metrics measure annotated-information recovery rather than precision over every field generated by the model.

5.2 Normalization and Matching

Keys are normalized deterministically using Unicode NFKC normalization, case-folding, decorative-punctuation normalization, separator-to-whitespace mapping, and repeated-whitespace collapse. This removes superficial formatting variation but does *not* use fuzzy matching, stemming, synonym expansion, semantic similarity, or dataset-specific aliases. Values use normalized text and whitespace tokenization for token-level scoring.

Keys match by exact equality after normalization. Duplicate occurrences are preserved and predictions cannot be reused. For each ground-truth pair, Value F1 is computed against unused predictions with the same normalized key; when multiple candidates exist, the one maximizing token-level F1 is selected and marked used. If none exists, the pair receives Value F1 zero.

5.3 Evaluation Metrics

We report three complementary metrics: Key Recall (KR), Exact Match (EM), and Value F1 (VF1).

Key Recall (KR): Let

$$K = \{k_i\}_{i=1}^{N}$$

denote the collection of normalized ground-truth key occurrences, and let $\hat{K}$ denote the corresponding collection of normalized predicted keys. Accounting for duplicate occurrences, Key Recall is

$$\mathrm{KR} = \frac{\sum_k \min\big(c_K(k), c_{\hat{K}}(k)\big)}{N},$$

where $c_K(k)$ and $c_{\hat{K}}(k)$ denote the number of occurrences of normalized key k in the ground truth and prediction, respectively.

Exact Match (EM): A ground-truth pair receives exact-match credit only when an unused prediction has both the same normalized key and an exactly matching normalized value. Each prediction can be matched at most once. Document-level EM is

$$\mathrm{EM} = \frac{N_{\mathrm{exact}}}{N},$$

where N_{exact} is the number of ground-truth pairs for which such an exact unused prediction exists.

Value F1 (VF1): Value F1 measures token-level overlap between a predicted value and its reference value. Let $c_{\mathrm{pred}}(t)$ and $c_{\mathrm{gt}}(t)$ denote the multiplicities of token t in the predicted and ground-truth values. Token overlap is

$$O = \sum_t \min(c_{\mathrm{pred}}(t), c_{\mathrm{gt}}(t)) \,.$$

Token-level precision and recall are

$$P = \frac{O}{|V_{\mathrm{pred}}|}, \qquad R = \frac{O}{|V_{\mathrm{gt}}|},$$

where $|V_{\mathrm{pred}}|$ and $|V_{\mathrm{gt}}|$ denote the numbers of tokens in the predicted and reference values, including repeated tokens. The corresponding score is

$$F_1 = \frac{2PR}{P+R},$$

with $F_1 = 0$ when there is no token overlap.

For every ground-truth pair, an unused prediction with the same normalized key is selected to maximize token-level F1. If no corresponding prediction exists, that pair receives $F_{1,i} = 0$. Document-level Value F1 is therefore

$$\mathrm{VF1} = \frac{1}{N} \sum_{i=1}^{N} F_{1,i}.$$

Importantly, the denominator is the total number of annotated ground-truth pairs rather than only the subset of recovered keys. Consequently, missing fields remain in the denominator and receive zero credit.

Thus, KR measures annotated-field recovery, EM complete KVP correctness, and VF1 partial value recovery while retaining missing annotated pairs in the denominator.

5.4 Macro Averaging

All reported dataset-level scores are macro-averaged over evaluable documents. For metric m,

$$\mathrm{Macro}(m) = \frac{1}{M} \sum_{d=1}^{M} m_d,$$

where M denotes the number of evaluable documents and m_d is the document-level metric.

Macro averaging gives each evaluable document equal weight regardless of its number of annotated KVPs; documents with no evaluable ground-truth KVPs are excluded before aggregation.

5.5 Evaluation Tracks

Gold inputs use the deterministic annotation-derived representations in Section 4.6; OCR inputs use PaddleOCR, EasyOCR, and Tesseract. Comparisons hold the model, prompt structure, parsing, and evaluation fixed so that the evaluated-document text source is the changing factor.

5.6 Design Choices and Scope of the Metrics

The protocol is intentionally reference-centered: fields absent from the annotation schema are not automatically counted as false positives. This is important for datasets such as SROIE, whose annotations cover only a subset of semantically meaningful receipt content. Consequently, KR, EM, and VF1 should not be interpreted as hallucination-frequency measures; unsupported or additional fields are examined descriptively.

Duplicate semantic keys are evaluated as occurrences rather than unique names, and predictions cannot be reused. Token-level VF1 uses token multiplicities rather than set intersection, preserving repeated tokens while assigning zero to unrecovered annotated pairs.

5.7 Reproducibility and Evaluation Validation

Raw generations, parsed predictions, and per-document scores are retained for recomputation and inspection. Metric sanity checks cover perfect extraction, correct keys with zero value overlap, missing keys, and mixtures of exact and missing pairs, verifying in particular that missing ground-truth pairs remain in the denominator.

For every document, the pipeline also records prompt length, input truncation, generated-token count, and output-limit events, allowing context-budget and generation-limit effects to be audited without post-hoc modification of predictions or scores.

Implementation code for prompt construction, model inference, output parsing, metric computation, and experiment execution is publicly available at https://github.com/zanvari/pixels-to-pairs. The repository supports reproduction and extension of the benchmark. Additional implementation details are provided in Appendix A.

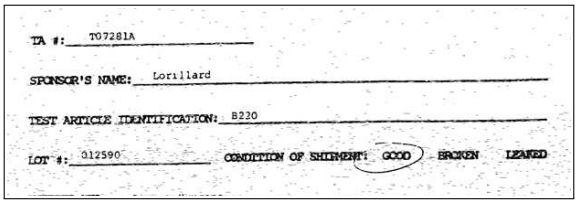
6 Results

This section reports the quantitative results on FUNSD, SROIE, and CORD under annotation-derived Gold text and OCR-derived inputs. We report macro-averaged Key Recall (KR), Exact Match (EM), and Value F1 (VF1) according to Section 5. Unless otherwise specified, comparisons across model families use zero-shot prompting; one- to three-shot experiments are reported separately for Qwen2.5–7B and LLaMA 3–8B.

6.1 Overall Findings

The results reveal three broad patterns. First, extraction difficulty varies substantially across document domains. Zero-shot Gold-text KR ranges from 0.2763–0.6332 across models on FUNSD, whereas several models approach complete key recovery on SROIE. Second, OCR-derived text generally produces substantial reductions in EM and VF1, although the magnitude of degradation

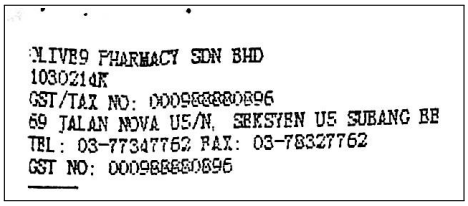
(a) Key-Label Corruption (FUNSD)



Gold excerpt	PaddleOCR excerpt
TA #: T07281A	[TA # label omitted] T07281A
SPONSOR'S NAME: Lorillard	SPONSOR'S NAME: LORIllard
TEST ARTICLE IDENTIFICATION:	TEST ARTTCKE IDENTIFICATION: :
B220	B220
LOT #: 012590	LoT #: 012590
CONDITION OF SHIPMENT:	CONDTTION OF SHIMENT:
GOOD BROKEN LEAKED	OODEROKEN IZAFED

Extraction impact: Qwen omits `TA #`, reproduces the corrupted sponsor and condition values, and preserves `012590` under the OCR-corrupted `LoT #` label.

(b) Character-Level Value Corruption (SROIE)



Gold excerpt	PaddleOCR excerpt
OLIVE9 PHARMACY SDN BHD	YLIVE9 PHARMACY SDN BHD
1030214K	1030214K
GST/TAX NO: 000988880896	GST/TAX NO: 000988880896
69 JALAN NOVA U5/N, SEKSYEN U5	69 JALAN NOVA US/N. EBESYEN
SUBANG BE	US SUBANG BB
TEL: 03-77347762 FAX:	TBL: 03-77347762 FAX:
03-78327762	03-78327762
GST NO: 000988880896	GST NO: 000988880896

Extraction impact: Qwen preserves the relevant field structure but propagates the OCR-corrupted company and address values.

(c) Cascading Value Misalignment (CORD)



Gold excerpt	PaddleOCR excerpt
1 | Lemon Tea (L.) | 25.000	25.000 | 1 | Lemon Tea (L.) |
1 | Mike & Ike Original |	28.000
28.000	Mike & Ike Original | 28.000
1 | Mike & Berry | 28.000	Mike & Berry | 7.000
1 | Roti Keju Manis. | 7.000	Roti Keju Manis. | 88.000 |
TOTAL | 88,000	TOTAL

Extraction impact: The displaced value sequence propagates across the item list, corrupting three item values and the total formatting.

(d) OCR-Induced Error Amplification (SROIE)



Gold excerpt	PaddleOCR excerpt
SECURITY & OA TRADING	SECURITY &OA TRADING:
GST NO: 000345100288	GST NO: 000345100288
DATE: 7/6/2016	Date: 7/0z018
TIME: 2:47:04 PM	Time: 2:47:04 PM

Extraction impact: Gold-text inference returns `date` → `7/6/2016`. From the malformed PaddleOCR string `7/0z018`, Qwen instead returns `date` → `7/02/2018`, converting the OCR error into a plausible-looking but incorrect date.

Fig. 3 Four qualitative error-propagation pathways for zero-shot Qwen2.5–7B under PaddleOCR-derived text: (a) **key-label loss and corruption** in FUNSD [12] document 87137840; (b) **character-level value corruption** in SROIE [10] document X51006328937; (c) **reading-order displacement with cascading value misalignment** in CORD [21] document receipt_00045; and (d) **OCR-induced error amplification** in SROIE document X51006619784. Each panel presents a focused document crop, the corresponding Gold and PaddleOCR excerpts, and the downstream extraction impact. Red marks the relevant OCR-induced changes.

depends on the dataset, OCR engine, and model. Third, in-context demonstrations can produce large gains for some model–dataset combinations, but these gains are neither monotonic in the number of demonstrations nor consistent across models.

These observations indicate that downstream KVP extraction depends jointly on document structure, upstream text quality, model family, and demonstration configuration rather than on any single factor in isolation.

6.2 OCR Quality Across Datasets

Before examining downstream extraction, Table 1 summarizes OCR quality using Character Error Rate (CER) and Word Error Rate (WER). Lower values indicate greater similarity to the corresponding reference text.

PaddleOCR has the lowest CER and WER on all three datasets. SROIE exhibits the lowest

Table 1 Character Error Rate (CER) and Word Error Rate (WER) for the evaluated OCR engines across the three benchmark datasets. Lower values indicate better OCR quality.

Dataset	OCR Engine	Mean CER	Mean WER
FUNSD	PaddleOCR	**0.4199**	**0.5675**
	EasyOCR	0.4974	0.7318
	Tesseract	0.5485	0.7197
SROIE	PaddleOCR	**0.3099**	**0.4577**
	EasyOCR	0.4145	0.7148
	Tesseract	0.4334	0.6615
CORD	PaddleOCR	**0.4844**	**0.6199**
	EasyOCR	0.5044	0.8519
	Tesseract	0.8031	0.9825

OCR error rates overall, whereas CORD is particularly challenging under Tesseract. CER and WER, however, do not uniquely determine downstream extraction performance because recognition errors differ in their location and semantic importance.

These are sequence-level error rates between each engine's native plain-text serialization and the corresponding annotation-derived reference, rather than isolated glyph-recognition error rates. Particularly on FUNSD, their magnitude reflects not only character substitutions and omissions but also differences in line grouping and reading order relative to the spatially reconstructed Gold text.

6.3 FUNSD Results

FUNSD is particularly challenging in the text-only setting because many question–answer relationships depend on two-dimensional layout. The corrected Gold-text representation described in Section 4.6 spatially reconstructs annotation text before flattening it into a sequence.

Table 2 reports zero-shot performance across Gold and OCR-derived text for the 47 evaluable FUNSD documents.

Qwen2.5–7B obtains the highest zero-shot KR, EM, and VF1 on the corrected Gold-text representation. Mistral–7B is the next strongest zero-shot model on all three metrics. The remaining gap between KR and EM across models indicates that detecting an annotated field does not guarantee recovery of its exact associated value.

Table 2 Zero-shot KVP extraction performance on **FUNSD** across Gold and OCR-derived text. Bold values indicate the best score within each text condition.

Text	Model	KR	EM	VF1
Gold	Gemma 2B	0.2763	0.2252	0.2404
	Gemma 7B	0.3990	0.3080	0.3433
	Mistral–7B	0.5850	0.4627	0.5102
	Qwen2.5–7B	**0.6332**	**0.5040**	**0.5519**
	LLaMA 3–8B	0.4384	0.3032	0.3594
PaddleOCR	Gemma 2B	0.3067	0.1352	0.1887
	Gemma 7B	0.5040	0.2211	0.3087
	Mistral–7B	**0.6403**	**0.2924**	**0.3690**
	Qwen2.5–7B	0.5703	0.2477	0.3532
	LLaMA 3–8B	0.4653	0.2072	0.2848
EasyOCR	Gemma 2B	0.1613	0.0296	0.0586
	Gemma 7B	0.3211	0.0574	0.1186
	Mistral–7B	**0.3897**	**0.0628**	**0.1433**
	Qwen2.5–7B	0.3841	0.0582	0.1405
	LLaMA 3–8B	0.2705	0.0394	0.0952
Tesseract	Gemma 2B	0.1760	0.0598	0.0941
	Gemma 7B	0.2008	0.0409	0.0712
	Mistral–7B	**0.3704**	0.1057	0.1775
	Qwen2.5–7B	0.3546	**0.1080**	**0.1784**
	LLaMA 3–8B	0.3002	0.0684	0.1230

6.3.1 Few-Shot Prompting on Gold Text

Figure 4(a) shows the zero- to three-shot trajectories under FUNSD Gold text; the corresponding numerical results are consolidated in Table 3.

The two models respond differently to demonstrations. LLaMA 3–8B improves substantially after adding demonstrations and reaches its highest scores at two shots. Qwen2.5–7B, in contrast, does not improve monotonically: zero-shot prompting gives its highest KR, while three-shot prompting gives slightly higher EM and VF1. These results show that increasing the number of demonstrations does not necessarily produce monotonic improvements.

6.3.2 OCR-Derived Inputs

The OCR-derived results in Table 2 show substantial differences among PaddleOCR, EasyOCR, and Tesseract.

PaddleOCR produces substantially stronger downstream results than EasyOCR and Tesseract for most configurations. Importantly, the relative ordering of models also changes across text conditions. Mistral–7B obtains the highest zero-shot KR, EM, and VF1 under PaddleOCR despite Qwen2.5–7B leading on Gold text. Under Tesseract, Mistral retains the highest KR, whereas Qwen2.5–7B has

the highest EM and VF1. Thus, model rankings under clean text do not necessarily transfer directly to OCR-derived inputs.

6.3.3 Few-Shot Prompting under PaddleOCR

Figure 4(b) shows the FUNSD PaddleOCR trajectories, and Table 3 reports the corresponding scores using the same clean training demonstrations as in the Gold-text experiment.

The trajectories show that additional demonstrations do not consistently improve extraction and can instead reduce it. Qwen2.5–7B obtains its highest PaddleOCR scores at three shots, but its two-shot VF1 falls below both one-shot and zero-shot performance. LLaMA 3–8B obtains its highest KR at three shots but its highest EM and VF1 at two shots. These results reinforce that demonstration count alone is not a reliable predictor of extraction quality.

6.4 Sensitivity to Demonstration Composition

Following the protocol in Section 4.8, we evaluated 20 frozen three-shot demonstration sets, five each under the Low, Representative, High, and Diverse strategies, on FUNSD Gold text with Qwen2.5–7B and LLaMA 3–8B, yielding 40 sensitivity configurations. The 60 training documents were not reused across sets, and all sets were frozen before inference without model-output or test-performance feedback.

A context audit found no truncation in the main FUNSD runs or in 38 of the 40 sensitivity configurations. Qwen High-4 truncated the target document in all 47 cases and was excluded from aggregation; LLaMA High-4 truncated 2/47 documents and was retained. Table 4 summarizes the valid configurations.

Low-complexity sets yield the lowest mean VF1 for both models: 0.5146 for Qwen2.5–7B and 0.5531 for LLaMA 3–8B. Representative sets produce the highest means (0.5785 and 0.5829, respectively), although LLaMA obtains a nearly identical mean under Diverse demonstrations (0.5823), confirming that increasing demonstration complexity is not monotonically beneficial.

Substantial within-strategy variation also remains. Across Representative sets, VF1 ranges from 0.5373 to 0.6176 for Qwen2.5–7B and from 0.5530 to 0.6321 for LLaMA 3–8B. Figure 5 visualizes the Qwen variation, demonstrating that conclusions based on a single demonstration set can depend materially on the selected examples.

6.5 SROIE Results

SROIE evaluates receipt extraction using a compact annotation schema. The dataset is substantially easier for key identification than FUNSD under the text-only setting, with several models approaching complete key recovery.

Gemma 7B obtains complete zero-shot KR, while Qwen2.5–7B obtains the highest EM and VF1. The difference between near-perfect key recovery and substantially lower EM shows that recognizing SROIE's compact field inventory is considerably easier than reproducing the corresponding values exactly.

6.5.1 Few-Shot Prompting

LLaMA 3–8B is particularly sensitive to demonstrations on SROIE: KR increases from 0.6542 at zero-shot to 0.9978 after one demonstration, with corresponding large gains in EM and VF1. Qwen2.5–7B begins with near-complete KR, so its few-shot gains are smaller and concentrated primarily in EM.

6.5.2 OCR-Derived Inputs

SROIE key recovery remains comparatively robust under OCR degradation, whereas EM and VF1 decline more substantially. For example, Table 5 shows that Qwen2.5–7B KR changes only from 0.9950 on Gold text to 0.9935 under PaddleOCR, while EM decreases from 0.6866 to 0.5742 and VF1 from 0.8930 to 0.8367. This pattern indicates that OCR corruption often preserves enough information to identify the field type while altering the value itself.

6.5.3 Few-Shot Prompting under PaddleOCR

Figure 4(d) shows the corresponding SROIE PaddleOCR trajectories; exact scores are reported in Table 3.

Few-shot prompting under PaddleOCR produces substantial gains for LLaMA 3–8B and more

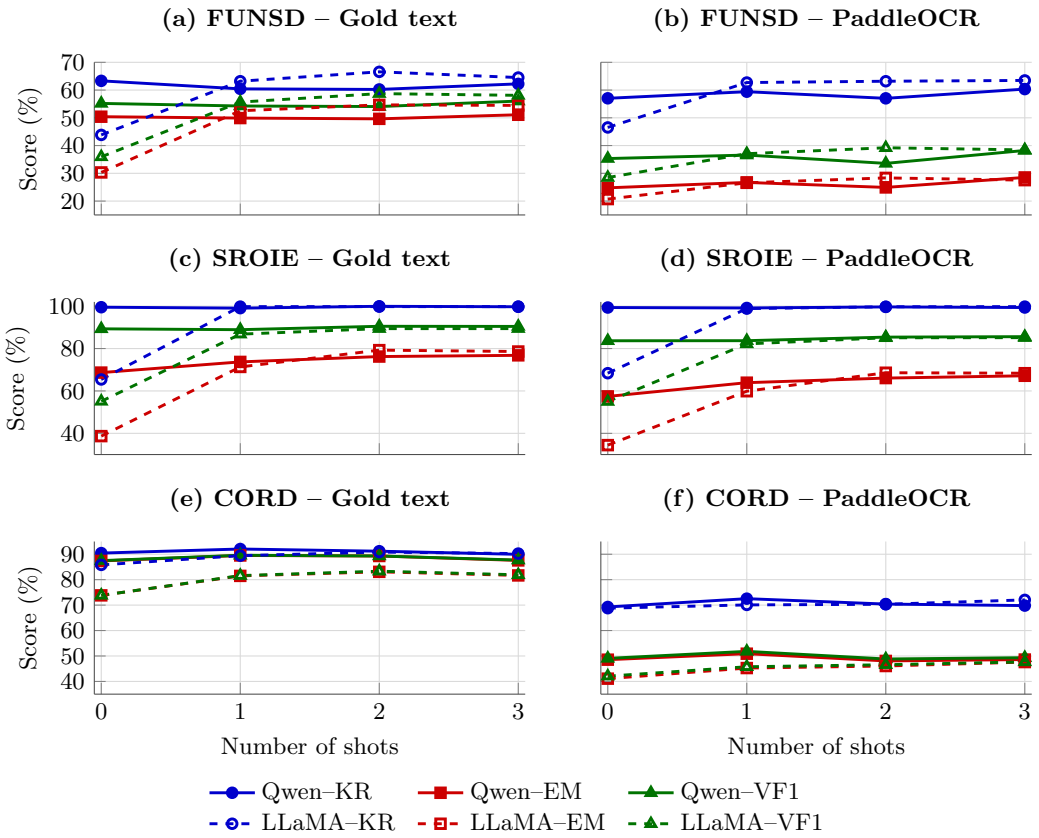


Fig. 4 Few-shot extraction trajectories for Qwen2.5–7B and LLaMA 3–8B across FUNSD, SROIE, and CORD under annotation-derived Gold text and PaddleOCR-derived text. Colors and marker shapes distinguish Key Recall (KR), Exact Match (EM), and Value F1 (VF1); solid filled markers denote Qwen2.5–7B, while dashed open markers denote LLaMA 3–8B. The trajectories demonstrate that the effect of additional demonstrations depends on the model, dataset, metric, and text condition, with no uniformly optimal demonstration count. Axis ranges are shared within each dataset row.

modest gains for Qwen2.5–7B. For Qwen2.5–7B, EM increases from 0.5742 at zero shot to 0.6715 at three shots, while VF1 increases from 0.8367 to 0.8561. LLaMA 3–8B shows a substantially larger response to demonstrations: KR increases from 0.6837 at zero shot to 0.9885 after one demonstration and reaches 0.9978 at three shots. Its highest EM is obtained at two shots, while its highest VF1 is obtained at three shots. As in the other experiments, additional demonstrations do not produce strictly monotonic improvements across all metrics.

6.6 CORD Results

CORD contains dense receipt information, repeated monetary values, and item-level fields. These characteristics create a different extraction challenge from both FUNSD and the compact SROIE schema.

Table 6 shows that Qwen2.5–7B obtains the highest zero-shot Gold-text scores, followed by LLaMA 3–8B. Compared with FUNSD, KR and EM are much closer for the strongest models, indicating that when CORD fields are recovered under clean text, their values can often also be extracted exactly.

6.6.1 Few-Shot Prompting

Figure 4(e) shows the CORD Gold-text trajectories; exact scores are reported in Table 3.

Both models initially benefit from demonstrations, but additional examples do not produce monotonic gains. Qwen2.5–7B reaches its highest scores at one shot and then declines, while LLaMA 3–8B reaches its highest scores at two shots. This behavior is consistent with the FUNSD experiments in showing that demonstration count

Table 3 Few-shot KVP extraction performance across datasets under annotation-derived Gold text and PaddleOCR-derived text. Within each dataset, the same clean training demonstrations were used under both text conditions. Bold values indicate the best score for each model, text condition, and metric across the evaluated shot counts.

			Gold Text			PaddleOCR		
Dataset	Model	Shots	KR	EM	VF1	KR	EM	VF1
FUNSD	Qwen2.5–7B	0	**0.6332**	0.5040	0.5519	0.5703	0.2477	0.3532
		1	0.6043	0.4991	0.5426	0.5941	0.2672	0.3658
		2	0.6023	0.4963	0.5406	0.5702	0.2492	0.3358
		3	0.6223	**0.5110**	**0.5603**	**0.6035**	**0.2852**	**0.3824**
	LLaMA 3–8B	0	0.4384	0.3032	0.3594	0.4653	0.2072	0.2848
		1	0.6319	0.5254	0.5567	0.6271	0.2658	0.3707
		2	**0.6657**	**0.5467**	**0.5872**	0.6318	**0.2835**	**0.3921**
		3	0.6451	0.5447	0.5810	**0.6347**	0.2754	0.3843
SROIE	Qwen2.5–7B	0	0.9950	0.6866	0.8930	0.9935	0.5742	0.8367
		1	0.9906	0.7370	0.8890	0.9914	0.6390	0.8371
		2	**0.9993**	0.7622	**0.9050**	**0.9964**	0.6607	0.8539
		3	0.9971	**0.7680**	0.9044	0.9935	**0.6715**	**0.8561**
	LLaMA 3–8B	0	0.6542	0.3869	0.5510	0.6837	0.3444	0.5500
		1	**0.9978**	0.7133	0.8679	0.9885	0.5987	0.8219
		2	**0.9978**	**0.7925**	0.8939	0.9971	**0.6859**	0.8508
		3	**0.9978**	0.7860	**0.8940**	**0.9978**	0.6837	**0.8514**
CORD	Qwen2.5–7B	0	0.9045	0.8736	0.8748	0.6927	0.4855	0.4906
		1	**0.9205**	**0.8951**	**0.8959**	**0.7253**	**0.5090**	**0.5182**
		2	0.9119	0.8931	0.8931	0.7041	0.4806	0.4882
		3	0.8996	0.8757	0.8770	0.6981	0.4860	0.4933
	LLaMA 3–8B	0	0.8581	0.7383	0.7383	0.6879	0.4109	0.4207
		1	0.8943	0.8149	0.8160	0.7008	0.4526	0.4577
		2	**0.9085**	**0.8308**	**0.8333**	0.7033	0.4606	0.4657
		3	0.9025	0.8167	0.8191	**0.7205**	**0.4761**	**0.4768**

Table 4 Sensitivity of three-shot FUNSD Gold-text performance to demonstration composition. Each cell reports mean ± sample standard deviation, with the observed range shown underneath.

Model	**Strategy**	n	**KR**	**EM**	**VF1**
Qwen2.5-7B	Low	5	0.5790 ± 0.0219 [0.5514, 0.6119]	0.4659 ± 0.0242 [0.4272, 0.4906]	0.5146 ± 0.0197 [0.4845, 0.5399]
	Representative	5	0.6530 ± 0.0329 [0.6169, 0.6932]	0.5437 ± 0.0455 [0.4933, 0.5961]	0.5785 ± 0.0374 [0.5373, 0.6176]
	High	4	0.6301 ± 0.0075 [0.6229, 0.6395]	0.5072 ± 0.0242 [0.4754, 0.5276]	0.5506 ± 0.0161 [0.5297, 0.5674]
	Diverse	5	0.6190 ± 0.0204 [0.6001, 0.6458]	0.5031 ± 0.0202 [0.4847, 0.5341]	0.5436 ± 0.0140 [0.5318, 0.5595]
LLaMA-3-8B	Low	5	0.6130 ± 0.0302 [0.5784, 0.6558]	0.5168 ± 0.0227 [0.4928, 0.5440]	0.5531 ± 0.0203 [0.5313, 0.5824]
	Representative	5	0.6578 ± 0.0315 [0.6254, 0.7095]	0.5459 ± 0.0349 [0.5069, 0.6000]	0.5829 ± 0.0297 [0.5530, 0.6321]
	High	5	0.6410 ± 0.0397 [0.5799, 0.6738]	0.5343 ± 0.0344 [0.4861, 0.5723]	0.5700 ± 0.0330 [0.5224, 0.6044]
	Diverse	5	0.6542 ± 0.0184 [0.6214, 0.6639]	0.5436 ± 0.0233 [0.5109, 0.5669]	0.5823 ± 0.0212 [0.5466, 0.6004]

alone is insufficient to characterize few-shot behavior.

6.6.2 OCR-Derived Inputs

As shown in Table 6, CORD exhibits substantial OCR sensitivity. For Qwen2.5–7B, VF1 decreases from 0.8748 on Gold text to 0.4906 under PaddleOCR, 0.2773 under EasyOCR, and 0.1488 under Tesseract. The deterioration is especially severe under Tesseract, consistent with its high CORD CER and WER. Model ordering is again metric-dependent: under EasyOCR, LLaMA 3–8B has slightly higher KR than Qwen2.5–7B, whereas Qwen2.5–7B retains higher EM and VF1.

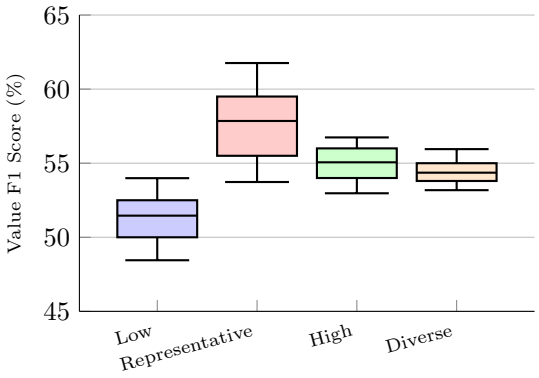


Fig. 5 Confirmatory Sensitivity Analysis: Variability in 3-shot Value F1 performance across demonstration complexity strategies (Qwen2.5-7B on FUNSD Gold). The observed ranges, for example, 53.73% to 61.76% within the Representative strategy, demonstrate substantial sensitivity to demonstration composition even when demonstration count and selection strategy are fixed.

Table 5 Zero-shot KVP extraction performance on **SROIE** across Gold and OCR-derived text. Bold values indicate the best score within each text condition.

Text	Model	KR	EM	VF1
Gold	Gemma 2B	0.8393	0.4107	0.5824
	Gemma 7B	**1.0000**	0.5778	0.8479
	Mistral–7B	0.9942	0.5836	0.8408
	Qwen2.5–7B	0.9950	**0.6866**	**0.8930**
	LLaMA 3–8B	0.6542	0.3869	0.5510
PaddleOCR	Gemma 2B	0.8624	0.3847	0.5735
	Gemma 7B	**0.9971**	0.5173	0.8067
	Mistral–7B	0.9870	0.4834	0.7824
	Qwen2.5–7B	0.9935	**0.5742**	**0.8367**
	LLaMA 3–8B	0.6837	0.3444	0.5500
EasyOCR	Gemma 2B	0.7976	0.1988	0.3351
	Gemma 7B	0.9856	0.3084	0.5621
	Mistral–7B	0.9777	0.3372	0.6161
	Qwen2.5–7B	**0.9870**	**0.4236**	**0.6568**
	LLaMA 3–8B	0.9157	0.3501	0.5977
Tesseract	Gemma 2B	0.8235	0.2363	0.3991
	Gemma 7B	**0.9510**	0.3646	0.6410
	Mistral–7B	0.9503	0.3653	0.6373
	Qwen2.5–7B	0.9388	**0.4330**	**0.6742**
	LLaMA 3–8B	0.7514	0.2961	0.5104

6.6.3 Few-Shot Prompting under PaddleOCR

Figure 4(f) shows the CORD PaddleOCR trajectories; exact scores are reported in Table 3.

Under CORD PaddleOCR, Qwen2.5–7B reaches its highest scores with one demonstration and then declines, whereas LLaMA 3–8B improves across the evaluated shot counts. The contrast provides another example of model-specific responses to in-context demonstrations.

Table 6 Zero-shot KVP extraction performance on **CORD** across Gold and OCR-derived text. Bold values indicate the best score within each text condition.

Text	Model	KR	EM	VF1
Gold	Gemma 2B	0.3140	0.2673	0.2682
	Gemma 7B	0.7022	0.6121	0.6121
	Mistral–7B	0.7820	0.6327	0.6361
	Qwen2.5–7B	**0.9045**	**0.8736**	**0.8748**
	LLaMA 3–8B	0.8581	0.7383	0.7383
PaddleOCR	Gemma 2B	0.2179	0.1385	0.1385
	Gemma 7B	0.5810	0.3486	0.3571
	Mistral–7B	0.5885	0.3357	0.3401
	Qwen2.5–7B	**0.6927**	**0.4855**	**0.4906**
	LLaMA 3–8B	0.6879	0.4109	0.4207
EasyOCR	Gemma 2B	0.1310	0.0576	0.0584
	Gemma 7B	0.4477	0.1737	0.1809
	Mistral–7B	0.4863	0.2087	0.2104
	Qwen2.5–7B	0.5176	**0.2698**	**0.2773**
	LLaMA 3–8B	**0.5206**	0.1888	0.1967
Tesseract	Gemma 2B	0.1225	0.0413	0.0413
	Gemma 7B	0.1972	0.0704	0.0734
	Mistral–7B	**0.2917**	0.0857	0.0884
	Qwen2.5–7B	0.2733	**0.1469**	**0.1488**
	LLaMA 3–8B	0.2777	0.0919	0.0998

6.7 Cross-Dataset Comparison

The three datasets expose different failure regimes. FUNSD remains difficult even with corrected Gold text because flattening a form into a text sequence removes direct spatial association cues. SROIE, with its compact field inventory, permits near-complete key recovery for several models, although exact value extraction remains more sensitive to OCR corruption. CORD combines dense receipt content with repeated numeric values and exhibits particularly large degradation as OCR quality decreases.

The results also show that OCR robustness cannot be summarized by a universal ordering of models. A model that performs best under Gold text may not retain the same relative position under OCR-derived input, and differences depend on whether KR, EM, or VF1 is considered. Similarly, few-shot prompting is not uniformly beneficial: the magnitude and direction of the effect depend on the model, dataset, number of demonstrations, and, as the FUNSD sensitivity experiment shows, the composition of those demonstrations.

Across the three datasets, three OCR engines, and three evaluation metrics, the top-ranked zero-shot model under OCR-derived input differs from the corresponding Gold-text leader in 10 of 27 comparisons.

Figure 6 visualizes this dataset-dependent OCR degradation for Qwen2.5–7B.

6.8 Observed Error Patterns

During inspection of saved model outputs and evaluation artifacts, we observed several recurring error patterns across datasets and text conditions. This inspection was descriptive rather than a separately sampled or quantitatively coded qualitative study. Observed patterns included key–value misalignment, OCR-induced numeric corruption, fragmented values, unsupported or unannotated fields, and structured-output failures.

Key–value misalignment is particularly relevant on FUNSD, where flattening two-dimensional forms removes direct spatial association cues. On SROIE and CORD, OCR corruption of dates, addresses, and monetary values can preserve the general field identity while preventing exact value recovery, consistent with the larger OCR-induced reductions in EM and VF1 than in KR for several configurations. Repeated monetary values in receipt documents also create ambiguity when multiple predictions share the same normalized key.

Additional generated fields require careful interpretation because the quantitative evaluation is reference-centered. Some additional fields may represent unsupported generations, whereas others may correspond to semantically meaningful document content that is not included in a dataset's annotation schema. Accordingly, KR, EM, and VF1 are not interpreted as direct measurements of hallucination frequency.

6.9 Comparison with Prior Document Understanding Approaches

End-to-end VLMs and layout-aware models are not used as numerical baselines because they consume pixels, layout coordinates, or both and often employ task-specific training and entity-level metrics. Our protocol instead varies only the text representation supplied to fixed text-only LLM inference. Combining their published scores with our document-level KR, EM, and VF1 would conflate modality, supervision, target representation, and evaluation. Their results are therefore complementary and require a separate multimodal protocol.

7 Discussion

The experiments show that text-only KVP extraction depends jointly on document structure, upstream text quality, model behavior, and prompting. The following sections discuss the implications of these interactions for OCR-based document understanding.

7.1 Semantic Capability, OCR Quality, and Document Structure

Across many model–dataset combinations, OCR affects exact value recovery more strongly than field identification. For example, models may continue to recognize that a receipt contains a `TOTAL` or `DATE` field while OCR corruption changes one or more characters in the associated value. Such an output can preserve key recovery while reducing Exact Match and Value F1. This distinction is important because apparently robust field detection does not necessarily imply reliable extraction of the information required by a downstream application.

The relationship between OCR quality and extraction performance is nevertheless not completely uniform. Lower CER and WER generally correspond to stronger downstream results at the dataset level, but the same aggregate OCR error rate can conceal errors of very different semantic importance. Corrupting a character in a monetary amount, date, or identifier may be more consequential for KVP extraction than several errors in unrelated document text. OCR robustness should therefore be understood in terms of both the amount and the location of transcription error.

Document structure further mediates this relationship. FUNSD relies heavily on spatial associations between questions and answers. Although spatially reconstructing its annotation text provides a substantially more appropriate clean-text representation than annotation-array ordering, the final LLM input is still linear text and does not explicitly encode two-dimensional relationships. SROIE, by contrast, has a compact field inventory

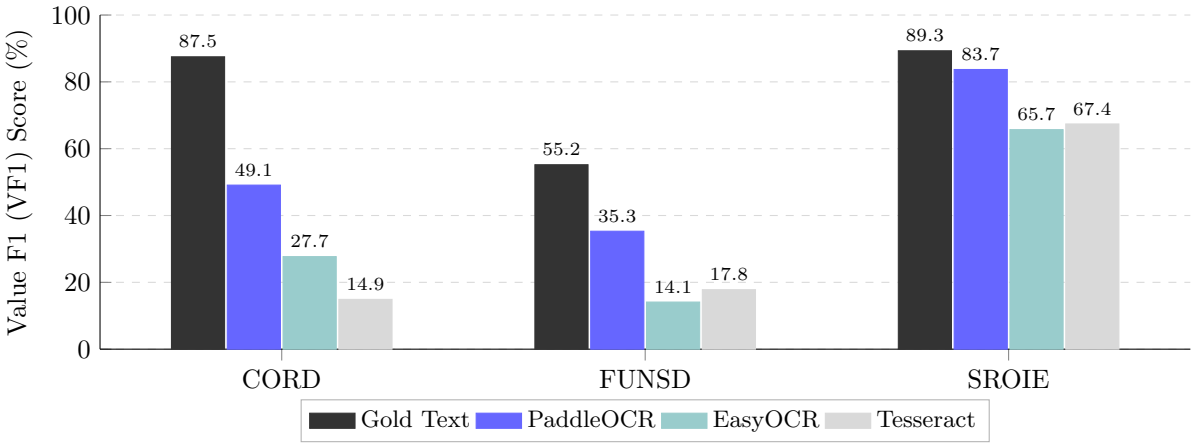


Fig. 6 Zero-shot Value F1 for Qwen2.5–7B across Gold and OCR-derived inputs. The magnitude of degradation varies substantially across datasets and OCR engines.

and relatively regular receipt structure, allowing high key recovery even when values are partially corrupted. CORD combines dense receipt content, repeated numeric values, and item-level information and exhibits particularly severe degradation under poorer OCR conditions. These differences demonstrate why robustness conclusions based on a single document dataset may not transfer directly to other document types.

7.2 Model Behavior under OCR Degradation

Model behavior is not uniform across datasets, metrics, and text conditions. Clean-text advantages do not necessarily persist after OCR degradation, and the relative ordering of models can change with both OCR condition and evaluation metric.

This behavior cautions against treating clean-text model rankings as direct predictors of OCR robustness. OCR changes the evidence supplied to the model, and different models respond differently to character corruption, altered tokenization, fragmented values, and disrupted reading order. Consequently, model selection for document extraction should be evaluated using textual inputs representative of the intended deployment pipeline rather than inferred solely from clean-text performance.

The results also do not support a simple claim that differences among models always converge as OCR quality deteriorates. Severe OCR corruption can constrain all models, as observed particularly on CORD under Tesseract, but relative behavior remains dataset- and metric-dependent. OCR quality therefore acts as an important constraint on downstream performance without eliminating differences among model families.

7.3 Few-Shot Prompting and Demonstration Sensitivity

Few-shot effects are strongly model-dependent and non-monotonic, showing that demonstration count alone does not determine extraction quality.

The confirmatory FUNSD sensitivity experiment further shows that demonstration *composition*, not only demonstration count, contributes to this variability. Holding the number of demonstrations fixed at three while changing their composition produces measurable differences in KR, EM, and VF1 for both Qwen2.5–7B and LLaMA 3–8B. Low-complexity demonstration sets yield lower mean VF1 than the representative strategy for both models, while increasing complexity further does not produce monotonic improvement. Variation also remains among independently sampled sets within the same strategy.

These results have two methodological implications. First, conclusions about few-shot performance based on a single demonstration set can partly reflect the particular examples selected rather than a general property of the model. Reporting the exact demonstrations is therefore important for reproducibility. Second, demonstration selection represents a meaningful experimental variable in document extraction. More demonstrations, or more complex demonstrations, are not

automatically preferable; their usefulness depends on the model and the extraction setting.

The context-feasibility audit further shows that demonstration selection affects not only extraction performance but also whether the complete target document can be accommodated under a fixed context budget.

7.4 Implications for Text-Only Document Understanding

The benchmark highlights a fundamental limitation of text-only document understanding. Once a document is converted into a linear text sequence, some spatial, visual, and hierarchical information is no longer explicitly available to the language model. Better prompting or a stronger language model may improve interpretation of the information that remains, but cannot reconstruct arbitrary evidence that has been removed or corrupted upstream.

This limitation is especially relevant to forms, tables, and visually complex receipts in which semantic relationships depend on relative position rather than text alone. At the same time, the results show that text-only LLMs can be effective when the textual representation is sufficiently informative, particularly because they enable flexible extraction without task-specific fine-tuning.

Text-only LLMs and layout-aware or multimodal document models should therefore be viewed as complementary design choices rather than directly interchangeable approaches. Layout-aware architectures preserve structural information, while instruction-tuned LLMs provide flexible semantic interpretation of textual content. Systems that combine reliable OCR, structural representations, and LLM-based reasoning may be better suited to heterogeneous real-world documents, particularly when layout and visual context carry information that cannot be preserved reliably in a flattened transcript.

7.5 Implications for Benchmarking and Deployment

The benchmark has three practical implications. First, annotation-derived text isolates downstream extraction behavior, whereas OCR-based evaluation remains necessary for deployment because aggregate CER and WER can conceal errors in task-critical values. Second, few-shot studies should report their demonstrations and assess sensitivity to their selection. Third, OCR, representation, prompting, and semantic extraction should be evaluated jointly: stronger LLMs cannot recover information lost upstream, and improved OCR cannot restore structural relationships removed during text serialization. Robustness should therefore be measured on the complete pipeline under target operating conditions.

Having established the specific vulnerabilities of text-only LLMs to upstream OCR degradation, particularly regarding layout displacement and value corruption, future work must explore targeted mitigation strategies. Specifically, investigating whether noisy in-context demonstrations, OCR-aware prompt engineering, or parameter-efficient fine-tuning (PEFT) can enable LLMs to dynamically reverse-engineer OCR structural errors remains a critical next step for decoupled Document AI pipelines.

8 Limitations

The controlled benchmark has several limitations.

8.1 Dataset and Evaluation Scope

The evaluated datasets cover forms and receipts from a limited set of document domains and languages. Although many target fields are represented using English labels, the source documents may contain non-English text. The benchmark was not designed to systematically evaluate cross-lingual generalization and does not cover the broader diversity of handwritten, multi-page, or structurally complex documents. Generalization beyond these settings therefore remains to be established.

Evaluation is also constrained by dataset annotation coverage. SROIE, for example, contains only four annotated fields (*company*, *date*, *address*, and *total*), although receipts may contain additional semantically valid information. Predictions for unannotated fields are therefore excluded rather than counted as false positives. Consequently, Key Recall, Exact Match, and Value F1 measure recovery of annotated key–value pairs rather than precision over the complete set of generated fields. Unsupported or additional predictions were observed during output inspection but were not quantified by the primary evaluation metrics.

Finally, the study relies on automatic evaluation. Formatting differences, annotation omissions, or semantically valid predictions outside the reference schema may not be fully reflected by the reported scores; human evaluation could provide complementary evidence of practical extraction quality.

8.2 Experimental Design

Results under OCR-derived inputs combine OCR and downstream extraction errors: substitutions, fragmentation, numeric corruption, reordering, and structural loss can all affect scores. Deterministic JSON-first parsing without semantic repair adds another failure source because malformed outputs may yield incomplete recovery despite partially correct content. We use a common prompt and decoding configuration for controlled comparison rather than per-model optimization.

Constrained- and open-schema prompting were not separately evaluated. SROIE lists its four fixed annotated fields, whereas FUNSD and CORD use open extraction because their keys vary by document. Schema listing may reduce unsupported outputs but suppress unlisted valid fields; open extraction offers greater flexibility but may generate additional fields. The reference-centered metrics do not quantify this precision–coverage trade-off.

8.3 Model Coverage

Resource constraints limited evaluation to instruction-tuned open-weight models of approximately 2B–8B parameters in a single-GPU environment. Whether the observed rank changes and non-monotonic few-shot behavior persist for substantially larger open-weight models or proprietary APIs remains open. Greater scale may improve instruction following and noise tolerance but cannot restore evidence absent from the serialized input. Task-specific fine-tuning, schema-guided or retrieval-augmented methods, and multimodal models are also outside scope because they introduce additional supervision or modalities.

9 Conclusion

This study presented a controlled benchmark of five open-weight instruction-tuned LLMs for text-only KVP extraction across FUNSD, SROIE, and CORD under annotation-derived and three OCR-derived text conditions. With prompting, inference, parsing, and evaluation fixed, OCR degradation affected exact and partial value recovery more strongly than field identification, and clean-text rankings did not reliably transfer to OCR-derived inputs. Few-shot prompting sometimes produced large gains but remained model-dependent, non-monotonic, and sensitive to demonstration composition.

These findings characterize text-only KVP extraction as a pipeline-level problem jointly governed by textual fidelity, document structure, model behavior, and demonstration choice. The released benchmark supports extension to additional domains, languages, OCR-aware methods, larger models, and multimodal systems.

A Implementation Details

All checkpoints used greedy decoding (`do_sample=False`), an 8,192-token input limit, up to 1,024 generated tokens, batch size one, and no task-specific fine-tuning. Runs were sequential on an NVIDIA A100 40 GB using Python 3.13.15, PyTorch 2.11.0, CUDA 12.8, and Transformers 5.16.1. OCR text remained uncorrected, and JSONL artifacts supported resumable execution.

Declarations

- **Funding:** No funding was received for conducting this study.
- **Competing Interests:** The author has no relevant financial or non-financial interests to disclose.
- **Author Contributions:** Zahra Anvari was the sole contributor to the study conception, design, execution, data analysis, and manuscript preparation.
- **Code Availability:** The custom code used in this study is publicly available at https://github.com/zanvari/pixels-to-pairs.

References

[1] AI@Meta (2024) The Llama 3 herd of models. arXiv preprint arXiv:2407.21783

[2] Anvari Z, Athitsos V (2021) A survey on deep learning based document image

enhancement. arXiv preprint https://doi.org/10.48550/arXiv.2112.02719, arXiv:2112.02719

[3] Anvari Z, Athitsos V (2025) Large language models for structured document understanding: A comprehensive survey of tasks, models, and challenges. TechRxiv

[4] Appalaraju S, et al (2021) DocFormer: End-to-end transformer for document understanding. In: ICCV

[5] Du Y, et al (2020) PP-OCR: A practical ultra lightweight OCR system. arXiv preprint arXiv:2009.09941

[6] Gemma Team, et al (2024) Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295

[7] Ha J, Haralick RM, Phillips IT (2000) Document analysis and understanding for paper-based forms. In: Document Analysis Systems Workshop

[8] Hu A, Xu H, Ye J, et al (2023) mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499

[9] Huang Y, et al (2022) LayoutLMv3: Unified text and image masking for document AI. arXiv preprint arXiv:2204.08387

[10] Huang Z, Gao L, Liu L, et al (2019) ICDAR 2019 competition on scanned receipt OCR and information extraction. In: ICDAR Workshops

[11] Jaided AI (2020) EasyOCR. https://github.com/JaidedAI/EasyOCR

[12] Jaume G, Ekenel HK, Thiran JP (2019) FUNSD: A dataset for form understanding in noisy scanned documents. In: ICDAR Workshops

[13] Jiang AQ, et al (2023) Mistral 7B. arXiv preprint arXiv:2310.06825

[14] Katti AR, et al (2018) Chargrid: Understanding 2d documents. In: EMNLP

[15] Kim G, et al (2022) OCR-free document understanding transformer. In: ECCV

[16] Lee J, et al (2021) FormNet: Structural encoding for form understanding. In: EMNLP

[17] Li Y, et al (2021) StructuralLM: Structural pre-training for form understanding. arXiv preprint arXiv:2105.11210

[18] Liu H, Li C, Wu Q, et al (2023) Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS)

[19] Luo C, et al (2024) LayoutLLM: Layout instruction tuning for document understanding. arXiv preprint arXiv:2404.05225

[20] OpenAI (2023) Gpt-4 technical report. arXiv preprint arXiv:2303.08774

[21] Park S, Shin M, Lee B, et al (2019) CORD: A consolidated receipt dataset for post-OCR parsing. In: NeurIPS Document Intelligence Workshop

[22] Qwen Team (2024) Qwen2.5 technical report. arXiv preprint arXiv:2412.15115

[23] Shilman M, Liang J, Haralick R (2005) Structured document image analysis using prior knowledge and heuristics. In: ICDAR

[24] Shpigel Nacson M, et al (2024) DocVLM: Efficient vision-language model for documents. arXiv preprint arXiv:2412.08746

[25] Smith R (2007) An overview of the Tesseract OCR engine. In: ICDAR

[26] Tang Z, et al (2022) UDOP: Unified document processing. arXiv preprint arXiv:2212.02623

[27] Wang X, et al (2023) InstructUIE: Multi-task instruction tuning for unified information extraction. arXiv preprint arXiv:2304.08085

[28] Wei J, Bosma M, Zhao VY, et al (2022) Fine-tuned language models are zero-shot learners. In: International Conference on Learning Representations

[29] Xu Y, et al (2020) LayoutLM: Pre-training of text and layout for document understanding. In: KDD

[30] Xu Y, et al (2021) LayoutLMv2: Multi-modal pre-training for visually-rich documents. In: ACL