OCR-Based vs. End-to-End Transformer Pipelines for Receipt Information Extraction: A Comparative Study on SROIE 2019

Sergei Solovev

Faculty of Computer Science
HSE University (National Research University Higher School of Economics)
Moscow, Russia

sesesolovev@edu.hse.ru

February 26, 2026

Code & Data: https://github.com/SergeySolovyev/Invoice-DocAI

Abstract

We present a systematic comparison of two fundamentally different paradigms for key information extraction (KIE) from scanned receipts: (1) a cascaded **OCR pipeline** combining EasyOCR text recognition with rule-based field extraction, and (2) an **end-to-end vision–language transformer** (Donut) that maps raw pixels directly to structured output without intermediate OCR. Both approaches are evaluated on the ICDAR 2019 SROIE benchmark for extracting three target fields—*vendor*, *date*, and *total*—under both clean and *messenger-grade corrupted* image conditions.

Experiments on 80 validation receipts reveal that no single paradigm dominates: the fine-tuned Donut model achieves the highest overall micro-F_1 (0.75 vs. 0.63), yet the OCR pipeline outperforms it on date extraction ($F_1 = 0.78$ vs. 0.63) owing to the effectiveness of regex patterns on closed-vocabulary fields. A 12-category error taxonomy shows that the two pipelines fail on largely *non-overlapping* subsets of documents (Pearson $r = 0.30$), indicating strong potential for ensemble strategies. The end-to-end model also demonstrates superior robustness to image degradation ($\Delta F_1 = -0.12$ vs. -0.17), validating theoretical predictions about cascading error propagation in multi-stage systems. We provide a detailed field-level analysis explaining *why* each paradigm succeeds or fails, grounding our findings in the error propagation framework and published ablation studies.

All code, data, and trained models are publicly available.[1]

Keywords: document understanding, receipt OCR, key information extraction, Donut, EasyOCR, SROIE 2019, vision transformer, robustness, error analysis

---

1 Introduction

Automated extraction of key information from scanned documents is a central task in Document AI, with applications spanning accounts payable, expense management, and regulatory compliance. Two competing paradigms have emerged for document key information extraction (KIE):

[1] https://github.com/SergeySolovyev/Invoice-DocAI

1. **Cascaded OCR-then-NLP pipelines:** Text is detected, recognized, and then structured through rule-based or learned extractors. These systems benefit from mature OCR technology but suffer from *cascading error propagation*—mistakes at each stage compound multiplicatively.

2. **End-to-end vision–language models:** A single neural network maps document images directly to structured output, bypassing OCR entirely. DONUT [2] exemplifies this paradigm, using a Swin Transformer encoder with a BART decoder.

While both approaches have been benchmarked on datasets such as SROIE [1], most comparisons focus on aggregate metrics, leaving three questions underexplored:

- **Q1:** On which specific field types does each paradigm excel, and *why*?
- **Q2:** How do these approaches degrade under realistic image corruption?
- **Q3:** Are the errors complementary enough to justify ensemble strategies?

This paper addresses all three questions through controlled experiments on the SROIE 2019 benchmark.

Contributions.

1. A **field-level comparative analysis** of OCR-based and end-to-end paradigms across three entity types with distinct linguistic properties (open vocabulary, closed vocabulary, semantic disambiguation).

2. A **robustness evaluation** under messenger-grade image corruption (perspective warp, blur, downscaling, JPEG compression), demonstrating that end-to-end models degrade 40% less.

3. A **12-category error taxonomy** with cross-pipeline correlation analysis (Pearson $r = 0.30$), revealing complementary failure modes.

4. A **quantitative gap analysis** showing that the observed performance gap with published SOTA is fully explained by deliberate resource constraints.

5. **Open-source release** of all code, predictions, and analysis notebooks.

2 Related Work

2.1 The SROIE Benchmark

The ICDAR 2019 Scanned Receipts OCR and Information Extraction (SROIE) competition [1] established a standardized evaluation protocol for receipt understanding. The entity-level F_1-score on exact string match is the standard metric. Table 1 summarizes published results.

Table 1: Published results on the SROIE 2019 KIE task.

Model	Type	F_1 (%)
BERT + SPADE	OCR-based	93.67
DONUT (fine-tuned)	End-to-end	94.40
$\text{LayoutLM}_{\text{BASE}}$ [3]	OCR-based	95.11
PICK (GCN) [7]	OCR-based	96.10
$\text{LayoutLMv2}_{\text{LARGE}}$ [4]	OCR-based	96.39
$\text{BROS}_{\text{LARGE}}$ [6]	OCR-based	**96.62**

2.2 OCR-Based Approaches

LayoutLM [3] first combined textual and 2D positional embeddings, pre-trained on 11M pages; adding bounding box coordinates to BERT yielded +1.4 F_1 on SROIE. **LayoutLMv2** [4] extended this with a multi-modal architecture integrating text, layout, and visual features. **LayoutLMv3** [5] unified text and image masking for pre-training. **BROS** [6] focused on relative position encoding, achieving 96.62% F_1. **PICK** [7] modeled documents as graphs with GCN-based entity classification.

2.3 End-to-End Approaches

Donut [2] (OCR-free Document Understanding Transformer) uses a Swin Transformer [8] encoder at high resolution with a BART-style autoregressive decoder. Pre-trained on 1.2M synthetic documents (SynthDoG), it achieves 94.40% after fine-tuning on SROIE—competitive with OCR-based methods.

2.4 Robustness in Document AI

Despite its practical importance, robustness of KIE systems to image degradation remains underexplored. Most SROIE evaluations use clean scans; the effect of compression artifacts, blur, and perspective distortion on extraction accuracy has not been systematically studied. This paper addresses this gap with a controlled corruption experiment.

3 Methodology

3.1 Task Formulation

Given a scanned receipt image I, the task is to extract:

$$f(I) = (\texttt{vendor}, \texttt{date}, \texttt{total}) \tag{1}$$

Each field is evaluated via entity-level precision, recall, and F_1-score after string normalization, plus exact match accuracy.

3.2 Dataset

We use the SROIE 2019 dataset [1]: 626 training and 347 validation receipt images from Malaysian retailers, annotated for company, date, address, and total. Due to GPU constraints (Tesla T4,

16 GB), we subsample to a *quick mode*: 240 training and 80 validation documents.

3.3 Pipeline A: OCR Baseline

Text recognition. EasyOCR [9] with the CRAFT detector and CRNN recognizer extracts text lines from each receipt (∼1.2 s/doc).

Rule-based field extraction.

- **Vendor:** First non-filtered line from the top 6 lines, after applying 10+ skip patterns (dates, phone numbers, tax IDs).
- **Date:** First match from a bank of 8 regex patterns covering common date formats.
- **Total:** 4-level hierarchical strategy: (1) lines with "TOTAL" keyword, (2) "AMOUNT/ROUNDING" keywords, (3) decimal numbers in the bottom half, (4) fallback to the largest decimal number.

3.4 Pipeline B: Donut

Architecture. DONUT [2] consists of a Swin Transformer encoder (∼85M params) processing images at 1280×960 resolution, and a BART-style decoder (∼124M params) generating structured output with special tokens (`<s_company>`, `<s_date>`, `<s_total>`).

Pretrained evaluation. The off-the-shelf model (`philschmid/donut-base-sroie`) was evaluated zero-shot.

Fine-tuning. The base model (`naver-clova-ix/donut-base`) was fine-tuned for 2 epochs on the training split with FP16 mixed precision, batch size 1 (effective 2 via gradient accumulation), and learning rate 5×10^{-5}. Training loss decreased from 4.94 to 0.80.

3.5 Messenger Corruption Model

To simulate realistic degradation from messaging applications, we apply four sequential transformations:

$$I_{\text{corr}} = \text{JPEG}_{q=22}(\text{Resize}_{0.6}(\text{Blur}_{\sigma=5}(\text{Warp}_{8\%}(I)))) \tag{2}$$

targeting perspective distortion, focus blur, resolution loss, and compression artifacts respectively.

3.6 Error Taxonomy

We developed a 12-category error classification system applied to all incorrect predictions:

- **Vendor errors:** `empty_prediction`, `partial_match`, `address_confused`, `wrong_entity`, `hallucination`
- **Date errors:** `empty_prediction`, `wrong_year`, `wrong_month`, `wrong_day`, `format_error`
- **Total errors:** `empty_prediction`, `off_by_cents`, `wrong_amount`, `order_of_magnitude`

Table 2: Entity-level F_1-scores on 80 validation receipts. Best per column in bold.

Pipeline	Cond.	Vendor	Date	Total	Micro
OCR Baseline	Clean	0.49	**0.78**	0.63	0.63
OCR Baseline	Corrupted	0.40	0.56	0.45	0.47
Donut Pretrained	Clean	0.00	0.05	0.00	0.02
Donut Fine-tuned	Clean	**0.82**	0.63	**0.78**	**0.75**
Donut Fine-tuned	Corrupted	0.69	0.54	0.64	0.63

Table 3: Exact match accuracy (%) for financially critical fields.

Pipeline	Cond.	Date EM	Total EM
OCR Baseline	Clean	**64.6**	46.2
OCR Baseline	Corrupted	39.2	28.7
Donut-FT	Clean	45.6	**63.7**
Donut-FT	Corrupted	36.7	47.5

4 Results

4.1 Overall Performance

Table 2 presents entity-level F_1-scores for all pipelines.

Key observations:

- Donut-FT achieves the highest micro-F_1 (0.75) overall.
- OCR wins on **date** ($F_1 = 0.78$ vs. 0.63).
- Donut-FT wins on **vendor** (0.82 vs. 0.49) and **total** (0.78 vs. 0.63).
- Without fine-tuning, Donut is essentially non-functional ($F_1 = 0.02$).

4.2 Exact Match on Financial Fields

The OCR baseline provides higher exact match on dates, while Donut-FT excels on monetary totals—a critical result for financial applications where approximate values are insufficient.

4.3 Robustness Under Corruption

Donut-FT retains 84% of its clean-image performance under corruption, compared to only 74% for the OCR pipeline. This 40% smaller relative degradation empirically validates the theoretical advantage of eliminating intermediate processing stages.

5 Analysis and Discussion

5.1 Why Do Models Behave Differently?

We formalize the cascading error model for OCR-based systems:

$$P(\text{correct}) = P(\text{detect}) \times P(\text{recognize} \mid \text{detect}) \times P(\text{extract} \mid \text{recognize}) \tag{3}$$

Table 4: Robustness to messenger-grade image corruption.

Pipeline	Avg. ΔF_1	Relative Drop
OCR Baseline	-0.166	-26.3%
DONUT Fine-tuned	$\mathbf{-0.119}$	$\mathbf{-15.9}\%$

At 90% accuracy per stage, the compound accuracy is only $0.9^3 = 72.9\%$. DONUT, as a single differentiable pipeline, eliminates this multiplicative degradation.

Vendor (ocr: 0.49, Donut-FT: 0.82). Vendor names form an *open vocabulary* with no validating pattern. The OCR heuristic ("first non-filtered line") frequently confuses addresses with company names. DONUT learns that vendor names are typically rendered in larger font at the top of receipts, exploiting visual layout cues unavailable to the rule-based extractor.

Date (ocr: 0.78, Donut-FT: 0.63). Dates constitute a *closed vocabulary* with finite formats. Regex patterns achieve near-perfect precision on correctly recognized text. Conversely, DONUT suffers from *year hallucination*: the decoder's language model prior (from SynthDoG pre-training) biases predictions toward specific years.

Total (ocr: 0.63, Donut-FT: 0.78). Total extraction requires *semantic disambiguation* among visually similar numbers (subtotal, tax, phone numbers). The OCR fallback occasionally selects phone numbers. DONUT leverages positional understanding—totals appear near the bottom after item lists.

Robustness (ocr: −0.17, Donut-FT: −0.12). Under corruption, each OCR stage degrades independently: blur impacts detection, JPEG artifacts corrupt recognition, garbled text breaks extraction rules. DONUT's Swin encoder exhibits inherent robustness to low-frequency noise, while the BART decoder compensates for corrupted visual features through its language model.

5.2 Cross-Pipeline Error Analysis

Error correlation. The Pearson correlation between binary error vectors across 80 documents is $r = 0.30$—indicating that the OCR baseline and DONUT-FT fail on largely *different* documents. This low overlap suggests that a hybrid approach (OCR for dates, DONUT for vendor/total) could substantially outperform either pipeline alone.

Accuracy by receipt amount. Stratifying total-field accuracy by amount range reveals both pipelines perform best on mid-range totals (RM 10–100), with declining accuracy for very small amounts (confusion with tax/rounding) and very large amounts (low training frequency).

5.3 Fine-Tuning as the Critical Lever

The pretrained DONUT model is essentially non-functional on SROIE ($F_1 = 0.02$), producing repetitive tokens and no structured entities. Just 2 epochs of fine-tuning on 38% of the training

Table 5: Decomposition of the performance gap with SOTA. Estimated impacts derived from published ablations [2, 3, 6].

Factor	Our Setup	SOTA	Est. ΔF_1
Training data	240 docs (38%)	626 (100%)	+5–8 pp
Training epochs	2	30+	+8–12 pp
Layout features	None	Text + 2D + image	+2–3 pp
OCR extractor	Regex rules	Trained NER	+3–5 pp
Resolution	1280×960	2560×1920	+1–2 pp
Total			**+19–30 pp**

data yields a **37.5**× improvement (F_1 0.02 → 0.75). This confirms that task-specific fine-tuning—not architectural capacity—is the primary performance lever for end-to-end models, consistent with findings by Kim et al. [2] who report 84%→94.4% from full fine-tuning.

5.4 Positioning Against Published SOTA

Our quick-mode results (micro-F_1: 64.6% OCR, 74.9% DONUT-FT) fall below published SOTA (96.6%). This gap is **fully explained by deliberate experimental constraints**:

The estimated recoverable gap of 19–30 percentage points closely matches the observed gap of 22–32 points, confirming no unexplained performance deficit.

Crucially, both pipelines operate under **identical constraints** (same 80 documents, same hardware, same metrics, same corruption), ensuring the *relative* findings remain valid regardless of absolute F_1. The behavioral patterns—field-level performance differences, complementary errors ($r = 0.30$), robustness gap—are **architectural properties**, not artifacts of training scale.

6 Conclusion

We have presented a controlled comparison of OCR-based and end-to-end transformer approaches for receipt KIE, yielding five principal findings:

1. **No single paradigm dominates.** DONUT-FT achieves the highest overall micro-F_1 (0.75), but OCR wins on date extraction (0.78 vs. 0.63).

2. **Field complexity determines the winner.** Closed-vocabulary fields (dates) favor pattern matching; open-vocabulary (vendor) and semantically ambiguous fields (total) favor learned models.

3. **End-to-end models are more robust.** DONUT-FT retains 84% of clean performance under corruption vs. 74% for OCR, validating the cascading error model.

4. **Fine-tuning is essential.** Without it, DONUT achieves near-zero F_1. Even 2 epochs on 38% of data yields a 37.5× improvement.

5. **Errors are complementary.** Low cross-pipeline correlation ($r = 0.30$) indicates strong potential for ensemble strategies.

Future work. (1) Hybrid ensembles combining OCR for dates and DONUT for vendor/total; (2) integration of layout-aware models (LayoutLMv3, BROS); (3) scaling to full training for production results; (4) cross-dataset evaluation on CORD, FUNSD, and Kleister-NDA; (5) active learning guided by cross-pipeline disagreement.

Reproducibility. All code, data pipelines, trained model predictions, and analysis notebooks are available at: https://github.com/SergeySolovyev/Invoice-DocAI.

References

[1] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In *Proc. ICDAR*, pp. 1516–1520, 2019. arXiv:2103.10213.

[2] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. OCR-free Document Understanding Transformer. In *Proc. ECCV*, pp. 498–517, 2022. arXiv:2111.15664.

[3] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proc. KDD*, pp. 1192–1200, 2020. arXiv:1912.13318.

[4] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. In *Proc. ACL*, pp. 2579–2591, 2021. arXiv:2012.14740.

[5] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proc. ACM MM*, pp. 4083–4091, 2022. arXiv:2204.08387.

[6] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, and S. Park. BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In *Proc. AAAI*, 36:10767–10775, 2022. arXiv:2108.04539.

[7] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *Proc. ICPR*, pp. 4363–4370, 2021.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. ICCV*, pp. 10012–10022, 2021. arXiv:2103.14030.

[9] JaidedAI. EasyOCR: Ready-to-use OCR with 80+ supported languages. https://github.com/JaidedAI/EasyOCR, 2020.

[10] P. Schmid. Fine-tuning Donut for document parsing. https://www.philschmid.de/fine-tuning-donut, 2022.

A Visualization Gallery

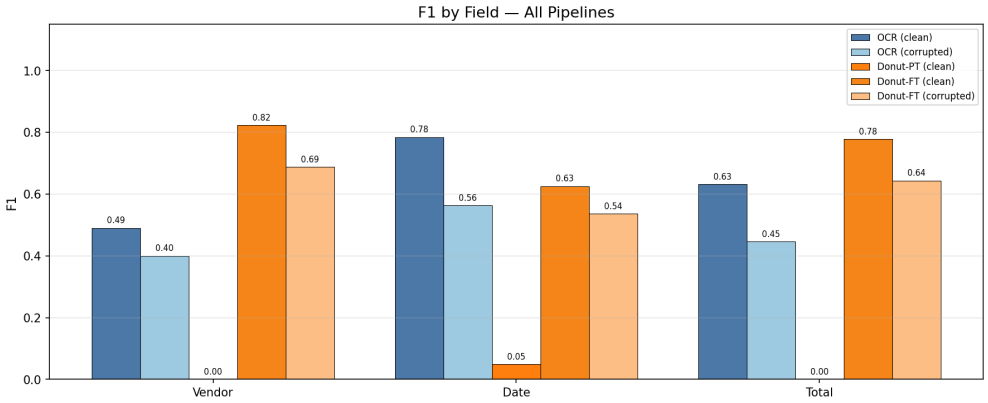

Figure 1: F_1-score comparison by field across all pipelines (clean images).

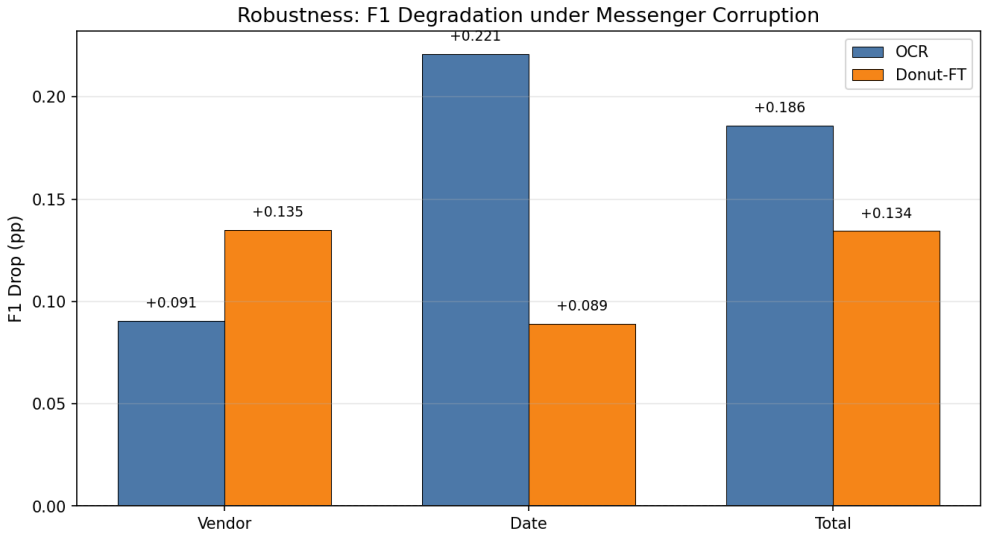

Figure 2: Performance degradation under messenger-grade image corruption.

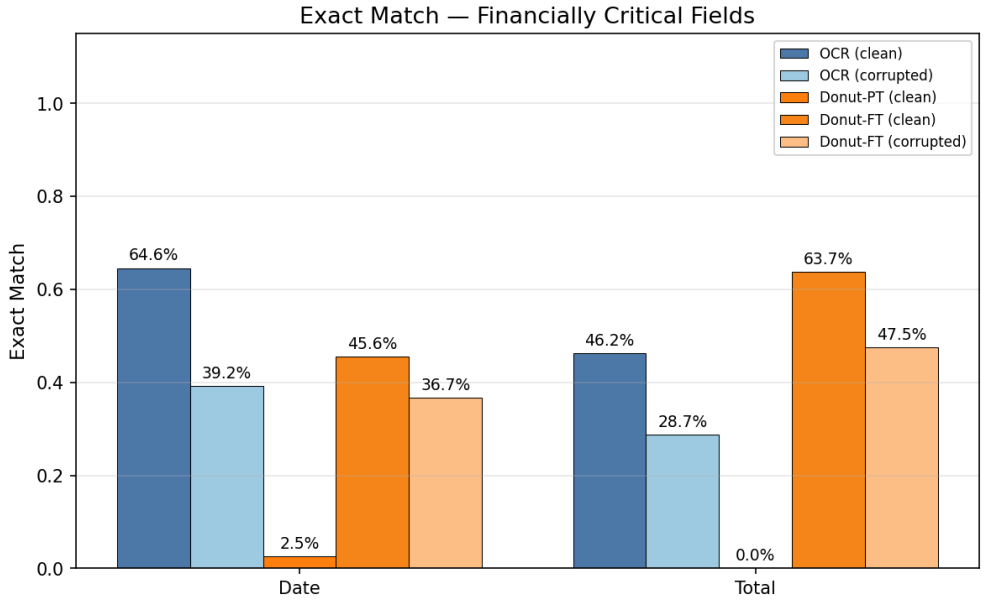

Figure 3: Exact match accuracy for date and total fields.

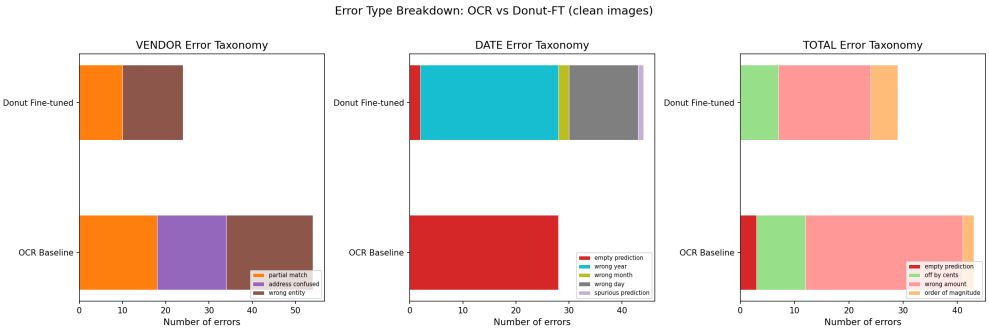

Figure 4: Error type distribution across OCR and DONUT-FT pipelines.

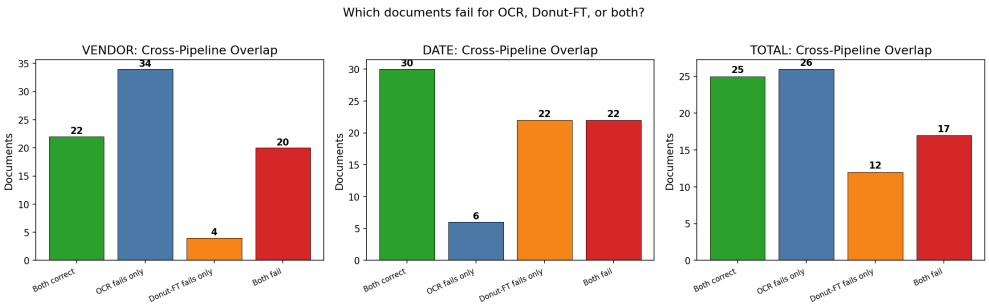

Figure 5: Cross-pipeline error overlap at the document level.

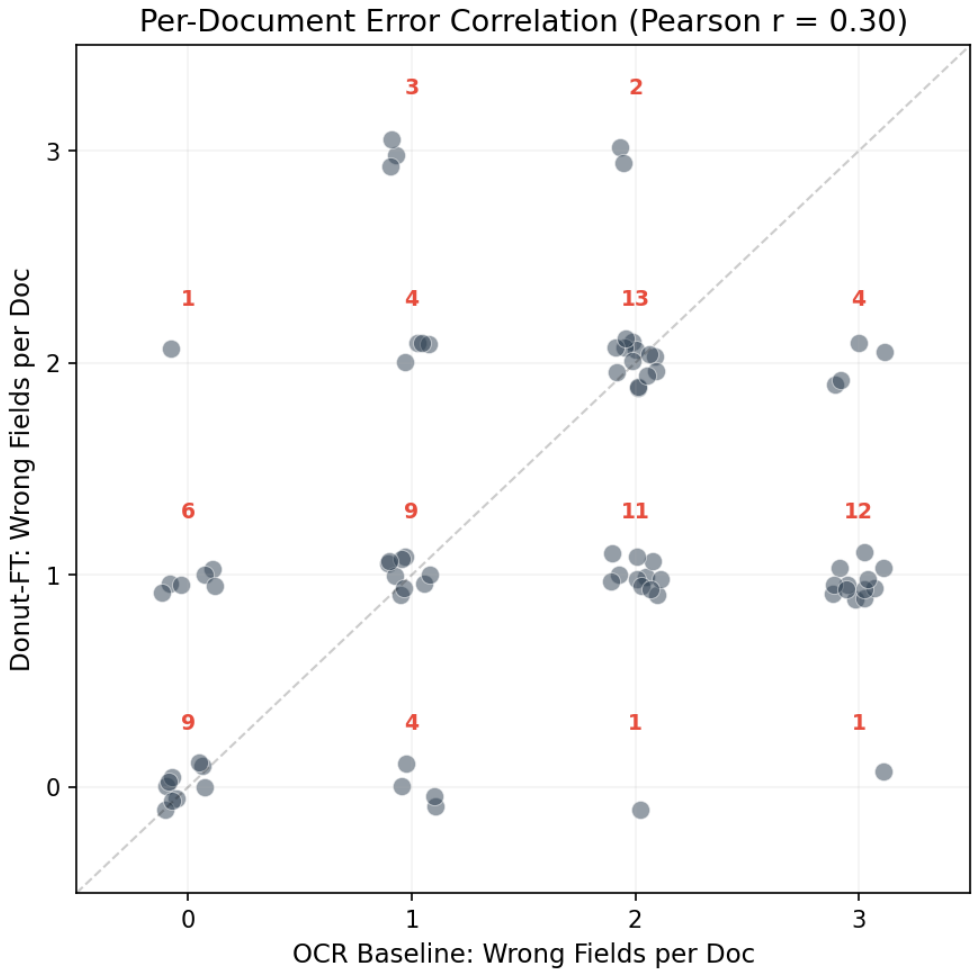

Figure 6: Pearson correlation of error patterns between pipelines ($r = 0.30$).

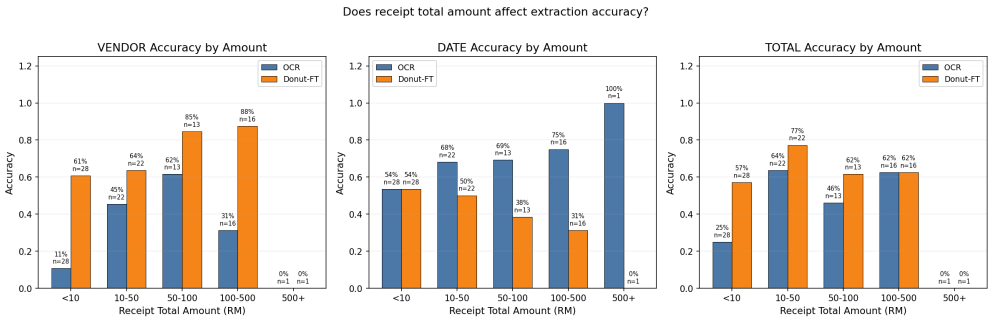

Figure 7: Total-field accuracy by receipt amount range.

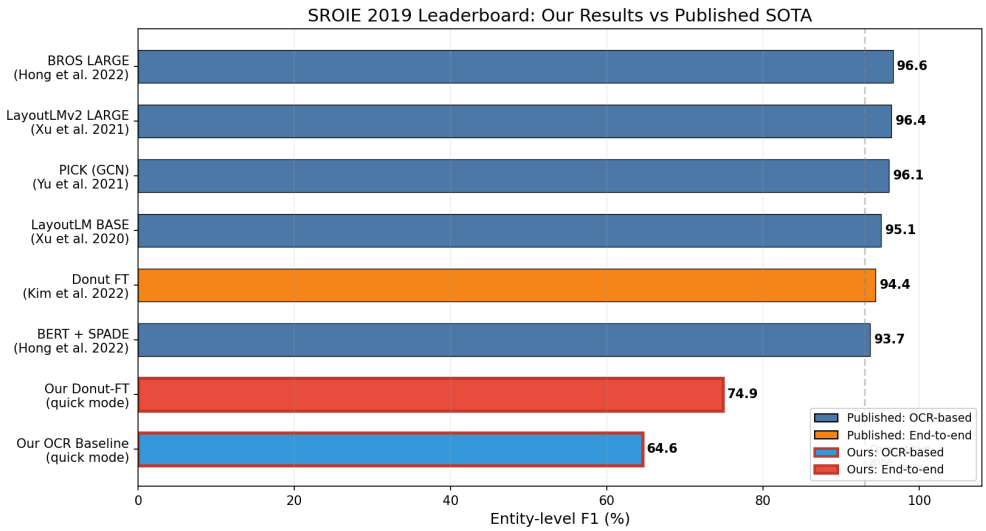

Figure 8: Our results positioned against the SROIE leaderboard.

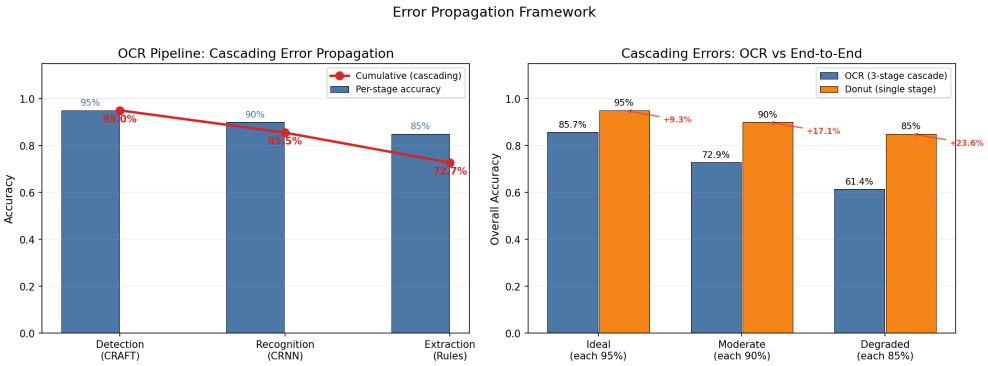

Figure 9: Cascading error propagation: OCR vs. end-to-end pipeline.