Reliability-Aware OCR-Free Extraction of Structured Data from Degraded Handwritten Medical Prescriptions

Abderrahman Kayouh
National School of Applied Sciences (ENSA),
Ibn Zohr University, Agadir, Morocco
Email: abderrahman.kayouh@edu.uiz.ac.ma

***Abstract*—We study OCR-free extraction of structured clinical fields (image→JSON) from Moroccan medical prescriptions using a fine-tuned Donut vision-language transformer (Swin encoder, mBART decoder). To confront the sim-to-real gap in the absence of annotated real data, we build a fully synthetic generator that injects realistic physical degradations (projective warping, paper folds, ink bleed, inhomogeneous lighting, sensor noise, stamps) and simulates handwriting through per-character stochastic jitter over five cursive fonts, with field bounding boxes co-transformed through the geometric warps. On 200 held-out images the model reaches 89.4% macro exact-match and 0.922 medication-list F1, and outperforms a classical modular OCR baseline by roughly 80 points (87.9% vs 6.1% macro EM over the textual fields). We further show that (i) string-based ontological normalization is nearly useless for the closed-vocabulary decoder, which already self-constrains its output, yet nearly doubles the same fields for the noisy modular baseline; (ii) decoder cross-attention is strongly grounded on the correct field ($9\times$ enrichment) *even when the prediction is wrong*, so attention location cannot detect hallucination; and (iii) the decoder's token confidence separates correct (0.994) from hallucinated (0.840) fields, enabling selective prediction that raises field precision from 88.0% to 98.6% at 83.5% coverage. All data are synthetic; the system is a research proof-of-concept, not a validated medical device.**

***Index Terms*—Document AI, OCR-free, vision-language transformer, Donut, uncertainty, selective prediction, synthetic data, medical prescriptions.**

I. Introduction

Extracting structured information from clinical documents is a core Document AI task with strong practical value and strong reliability constraints: in health, an unnoticed extraction error is worse than an explicit abstention. Classical pipelines chain text detection, OCR and post-hoc parsing, accumulating errors at each stage. OCR-free models such as Donut [1] instead map the raw image directly to a structured sequence, removing intermediate error propagation.

We extract nine fields (patient name, age, CIN, physician, specialty, city, date, medications, signature) from prescription images. Two constraints shape the work. First, real prescriptions contain sensitive personal data and cannot be freely collected; we therefore train entirely on *synthetic* data and study how far a realistic degradation model narrows the sim-to-real gap. Second, a deployable clinical extractor must *know when it does not know*. Our contributions are:

- A parametric synthetic generator combining physically-motivated degradations and per-character handwriting jitter, with automatically tracked, geometry-aware field bounding boxes.
- An honest multi-metric evaluation (EM, CER, WER, nested-list F1) of a fine-tuned Donut, a modular-baseline comparison, and evidence that residual errors are *semantic hallucinations*, not spelling noise.
- An interpretability result—attention is well grounded but grounding does *not* predict correctness—and a reliability result—decoder confidence does, enabling selective prediction (88.0%→98.6%).

II. Related Work

OCR-free document understanding. Donut [1] and Pix2Struct [2] generate structured text from images without an external OCR, using a vision-transformer encoder [3] and an autoregressive text decoder [4]. Layout-aware models [5] and OCR models such as TrOCR [6] remain strong in the modular paradigm. **Synthetic data and domain randomization.** Randomizing nuisance factors at training time [7] promotes invariance and is widely used when real annotations are scarce. **Reliability of generative decoders.** Selective prediction and the risk-coverage trade-off [8], [9] formalize abstention; we use token-level confidence as the selection score.

III. Methodology

A. Synthetic Prescription Generator

Each record is sampled from closed vocabularies (names, cities, specialties, publicly-listed over-the-counter drugs). Printed form labels use a sans font; handwritten values are drawn *glyph by glyph*, each glyph undergoing random rotation ($\pm 7^\circ$), scale, vertical drift and spacing, over five cursive fonts—approximating the intra- and inter-writer variability that fixed fonts cannot. We then apply physical degradations: a projective transform (homography), grid and elastic distortion (paper folds), morphological erosion/dilation (ink bleed/fade), low-frequency multiplicative lighting, additive Gaussian and intensity-dependent (shot) sensor noise, motion/Gaussian blur, and translucent stamps. Field bounding boxes are propagated as keypoints through the same geometric transforms so labels remain aligned with the warped image.

B. OCR-Free Model

We fine-tune `naver-clova-ix/donut-base`. The target JSON is serialized with 24 XML-like special tokens (e.g. `<s_cin_patient>`) added to the tokenizer, with the decoder embedding matrix resized accordingly. Images are processed at 960×720. We train for 10 epochs, learning rate 3×10^{-5}, effective batch size 4 (gradient accumulation), mixed precision. Padding is masked with `-100` in the loss; the target omits the leading BOS token to avoid a duplicated start symbol.

C. Ontological Normalization

Closed-vocabulary fields (city, specialty, drug name) are projected onto the nearest reference entry using a weighted string similarity (`WRatio`, Levenshtein-based). Fields without a controlled vocabulary (CIN, names, date) are left untouched.

D. Attention Grounding

For each generated token we average the decoder–encoder cross-attention over heads and layers, giving a weight over the 30×23 Swin patch grid, and aggregate it over a field's value tokens. Given the ground-truth box, we report the attention *enrichment* (in-box mass divided by box-area fraction) and the peak-in-box rate (one-patch tolerance).

E. Uncertainty and Selective Prediction

We take the decoder's per-token probability (softmax of the transition logits) and average it over a field's value tokens as the field confidence. Thresholding this confidence yields a selective predictor with a coverage/precision trade-off.

IV. Experimental Setup

We generate 1000 images (800 train / 200 validation, seed-fixed for reproducibility). Metrics: per-field exact match (EM), character and word error rate (CER/WER), and multiset precision/recall/F1 on the medication list.

V. Results

A. Field Extraction

Table I reports per-field performance. Macro EM is 89.4%; the medication list reaches F1$=0.922$ (P$=0.923$, R$=0.921$). The CIN field is the weakest (80.0%): a high-entropy random string with no linguistic context, every character must be read exactly. A clean-data variant reached 98.3% on the same fields; the $\sim$9-point drop quantifies the honesty of the degraded benchmark.

B. Comparison with a Modular Baseline

We compare against a classical pipeline: Otsu binarization, Tesseract OCR (French), and rule-based field parsing (Table II). The baseline collapses on the handwritten, degraded values (patient name 0.5%, CIN 0.5%), confirming that character-level OCR is ill-suited to this input, whereas the OCR-free model reaches 87.9% macro EM—a $\sim$80-point gap that justifies the end-to-end design. Notably, ontological normalization nearly doubles the closed-vocabulary fields of the baseline (specialty 11.5$\rightarrow$21.5, city 10.5$\rightarrow$20.0) because it repairs genuine OCR spelling noise—the exact opposite of its null effect on the self-constraining Donut decoder (Sec. III-C). Ontology is thus useful precisely where the recognizer emits out-of-vocabulary text.

TABLE I
PER-FIELD RESULTS ON 200 DEGRADED VALIDATION IMAGES.

Field	EM (%)	CER	WER
Patient name	90.0	0.048	0.064
Age	91.0	0.062	—
CIN	80.0	0.086	—
Physician	82.5	0.085	0.114
Specialty	91.0	0.073	0.125
City	92.0	0.128	—
Date	88.5	0.024	—
Signature (bool)	100.0	0.000	—
Macro EM	**89.4**		
Meds F1		**0.922**	

TABLE II
FIELD-LEVEL EXACT MATCH (%): MODULAR BASELINE VS. OURS.

Field	Tesseract	+ Ontology	Donut (ours)
Patient name	0.5	0.5	**90.0**
Age	4.5	4.5	**91.0**
CIN	0.5	0.5	**80.0**
Physician	12.0	12.0	**82.5**
Specialty	11.5	21.5	**91.0**
City	10.5	20.0	**92.0**
Date	3.0	3.0	**88.5**
Macro EM	6.1	8.9	**87.9**

C. Attention Is Grounded but Not a Reliability Signal

Fig. 1 overlays the aggregated cross-attention for three fields: it localizes precisely on the correct value. Quantitatively, attention concentrates on the true field at $\sim 9\times$ the uniform baseline (correct: 8.71$\times$, hallucinated: 9.18$\times$; peak-in-box 61.6% vs 71.2%). Crucially, grounding is as strong—or stronger—for wrong predictions: the model looks at the right region and still misreads it. Attention location therefore cannot detect hallucination.

D. Decoder Confidence Enables Selective Prediction

Mean field confidence is 0.994 for correct vs 0.840 for hallucinated fields. Fig. 2 shows the risk-coverage curve: abstaining on the 16.5% least-confident fields raises precision from 88.0% to 98.6%; at 90.8% coverage precision is already 95.2%. Confidence, not attention, is the operative reliability signal.

VI. Error Analysis

Two failure modes dominate. (1) *Semantic hallucination*: on small, low-context header fields (city, specialty) under strong degradation, the decoder falls back to frequent training values (a prior bias), producing valid-but-wrong outputs—which is exactly why string ontology cannot help. (2) *High-entropy reading*: the CIN, lacking redundancy, is the hardest

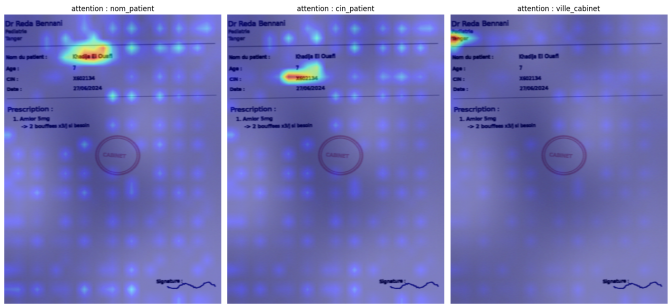

Fig. 1. Decoder cross-attention (red = high) for three fields on a correctly-read prescription; attention is grounded on each corresponding value.

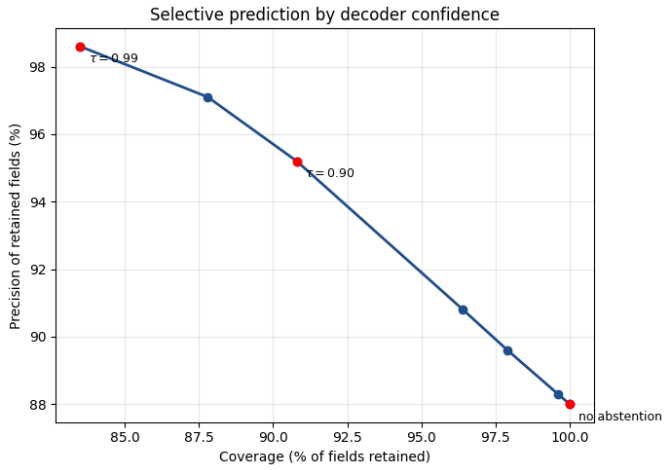

Fig. 2. Selective prediction by decoder confidence: precision of retained fields vs coverage.

field. Minor modes include field omission and occasional tag-boundary leakage.

VII. Limitations and Future Work

The evaluation is in-distribution: validation shares the generator's distribution, so absolute scores upper-bound real performance. Cursive fonts with jitter approximate but do not equal true handwriting; generative handwriting synthesis and fine-tuning on a small set of *real, consented, anonymized* prescriptions are natural next steps. A semantic drug ontology (brand→active ingredient/ATC) would add clinical interoperability independent of reading accuracy. Finally, a zero-shot vision-language model (e.g. Qwen2-VL, Florence-2) would provide a modern non-specialized baseline.

VIII. Ethics and Data Statement

All data are fully synthetic, generated by random combination of common Moroccan names, cities and publicly-listed over-the-counter drugs; no real patient or physician data were collected or used. This work is a research proof-of-concept and is not a validated medical device. Deployment on real prescriptions would require institutional authorization, CNDP compliance (Law 09-08), patient consent and anonymization.

IX. Conclusion

A Donut model trained purely on physically-degraded synthetic prescriptions extracts structured fields at 89.4% macro EM and 0.922 medication F1, and outperforms a modular OCR baseline by ∼80 points. Its errors are semantic hallucinations that string ontology cannot fix and attention grounding cannot flag, but that decoder confidence reliably detects—turning an 88.0% extractor into a 98.6%-precision one under selective prediction. Reliability, not raw accuracy, is where OCR-free clinical extraction must be engineered.

References

[1] 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. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[2] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

[3] 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. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 7871–7880.

[5] 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. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2020, pp. 1192–1200.

[6] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2023.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 23–30.

[8] R. El-Yaniv and Y. Wiener, "On the Foundations of Noise-free Selective Classification," *J. Mach. Learn. Res.*, vol. 11, pp. 1605–1641, 2010.

[9] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4878–4887.