RxLens: Multi-Agent LLM-powered Scan and Order for Pharmacy

Akshay Jagatap
Amazon
ajjagata@amazon.com

Srujana Merugu
Amazon
smerugu@amazon.com

Prakash Mandayam Comar
Amazon
prakasc@amazon.com

Abstract

Automated construction of shopping cart from medical prescriptions is a vital prerequisite for scaling up online pharmaceutical services in emerging markets due to the high prevalence of paper prescriptions that are challenging for customers to interpret. We present RxLens, a multi-step end-end Large Language Model (LLM)-based deployed solution for automated pharmacy cart construction comprising multiple steps: redaction of Personal Identifiable Information (PII), Optical Character Recognition (OCR), medication extraction, matching against the catalog, and bounding box detection for lineage. Our multi-step design leverages the synergy between retrieval and LLM-based generation to mitigate the vocabulary gaps in LLMs and fuzzy matching errors during retrieval. Empirical evaluation demonstrates that RxLens can yield up to 19% - 40% and 11% - 26% increase in Recall@3 relative to SOTA methods such as Medical Comprehend and vanilla retrieval augmentation of LLMs on handwritten and printed prescriptions respectively. We also explore LLM-based auto-evaluation as an alternative to costly manual annotations and observe a 76% - 100% match relative to human judgements on various tasks.

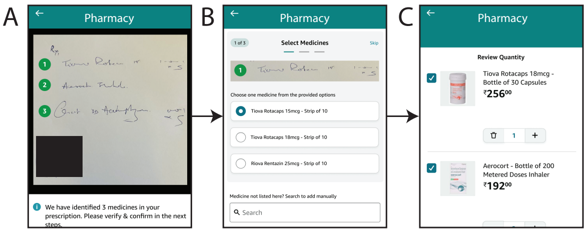
1 Introduction

Global adoption of online pharmacy services has surged in recent years, driven by demand for convenient, affordable access to medications. However, in emerging markets, paper prescriptions, which are typically unstructured, handwritten, and illegible, pose a major barrier for customers ordering medications online. Patients often report difficulties in deciphering doctors' handwriting accurately enough to use traditional e-commerce search. To mitigate the digitization errors and the consequent health risks, e-pharmacies offer "medicine dispensation" services where customers can upload prescriptions and receive cart-building assistance through either asynchronous digitization or direct pharmacist callbacks. While pharmacist calls provide better accuracy and capture specific needs like medication quantities and alternatives, they are costlier. Both approaches face scalability challenges due to the reliance on human pharmacists, resulting in long wait times and high cart abandonment. Hence, there is an urgent need for an automated, rapid, accurate, and scalable prescription digitization system to enable seamless online pharmacy ordering.

Building automated prescription-to-cart systems poses several key challenges. These span handling diverse layouts and handwriting styles, varying image quality and orientation, and region-specific medical terminology. Further, typos frequently cause confusion between similar drug names, making high accuracy critical for patient safety. A practical system must also secure patient PII while precisely mapping medications to the visual region on prescriptions. Lastly, the sensitive nature of prescriptions combined with expensive annotation effort leads to a significant scarcity of ground truth, complicating system development and evaluation.

Related Work. Current prescription digitization methods (Sharma et al., 2023; Guzman et al., 2020) follow a multi-step process: (a) optical character recognition, (b) medication extraction using custom-trained text and/or layout encoder models, and (c) matching extracted medications against a catalog. These methods perform poorly on non-US and handwritten prescriptions due to vocabulary gaps and limited training data. Studies on handwritten prescriptions (Gupta and Soeny, 2021; Davis and FACSM., 2008; Fajardo et al., 2019) have achieved limited success in identifying medicine names. Despite the broad success of recent foundational generative LLMs and multimodal approaches (Anthropic, 2023; McKinzie et al., 2024), their adoption for prescription digitization remains minimal. These models, trained pri-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 822–832
April 30, 2025 ©2025 Association for Computational Linguistics



[redacted]

We have identified 3 medicines in your prescription. Please verify & confirm in the next steps.

Choose one medicine from the provided options

● Tiova Rotacaps 15mcg - Strip of 10

○ Tiova Rotacaps 18mcg - Strip of 10

○ Riova Rentazin 25mcg - Strip of 10

Medicine not listed here? Search to add manually

Search

☑ Tiova Rotacaps 18mcg - Bottle of 30 Capsules ₹256⁰⁰

1

☑ Aerocort - Bottle of 200 Metered Doses Inhaler ₹192⁰⁰

Figure 1: Schematic view of automatic "scan and order" cart building from the prescription image.

marily on public datasets with limited handwritten documents and regional medical vocabulary, fail to achieve the desired accuracy when used directly or with vanilla retrieval augmentation, often due to hallucination. Further, the high LLM deployment costs (Sharir et al., 2020; Hoffmann et al., 2022) and PII concerns with third-party LLM APIs complicate their use in prescription digitization. Appendix A presents additional related work.

Contributions. We explore how to use LLMs (both multimodal and text-only) to develop an automated prescription-to-cart system. We investigate choices related to solution architecture, component-specific design, annotation scaling, and practical deployment, and present the below key contributions: 1) Building on existing methods, we propose RxLens, a modular LLM-based architecture comprising OCR, medication extraction, and matching against the catalog. This multi-step design leverages catalog-based retrieval augmentation to ensure medication validity. Within each step, we explore the benefits of LLMs and prompting strategies, focusing on the synergy between retrieval and generation. 2) We present solutions for handling practical system requirements, such as PII redaction before LLM invocation, medication-to-prescription region mapping, and latency optimization. 3) To address the lack of annotations and expensive labeling, we develop an LLM-based auto-evaluation approach using prompts that mimic human annotation (75.7% - 100% correlation). 4) Empirical evaluation shows RxLens achieves significant improvements (+19%-40% and +11%-26% Recall@3) over SOTA baselines like Medical Comprehend and vanilla LLM retrieval augmentation on handwritten and printed prescriptions, respectively.

2 Prescription Image Digitization

Formally, given a medicine catalog $\mathcal{A}$ [1], a prescription image P, and K, the max. number of suggestions per prescription item, the digitization process generates a list of s medication groups, $\hat{M}_{\mathcal{A}}(P) = \{g_1, \ldots, g_s\}$. Each group $g_i = (\mathbf{v_i}, \mathbf{a_i})$ includes a visual rectangular region of the prescription $\mathbf{v_i}$ and an ordered list of relevant medications $\mathbf{a_i} = \{a_{i1}, \ldots, a_{iK}\} \subset \mathcal{A}$. Let $M^*_{\mathcal{A}}(P) = \{g^*_1, \ldots, g^*_{s^*}\}$ denote the ideal cart with s^* groups where each group $g^*_i = (\mathbf{v}^*_i, \{a^*_{i1}\})$ contains the correct visual region and medication. Let $\rho : \{1, \cdots, s^*\} \mapsto \{1, \cdots, s\}$ map the medication groups in the ideal cart to the predicted ones [2]. The goal of digitization is to optimize the medication ranking and the visual region detection:

$$\max_{\hat{M}_{\mathcal{A}}(P)} \left(\sum_{i=1}^{s^*} L^{rank}(\mathbf{a}_i, \mathbf{a}_{\rho(i)}) + \lambda L^{visual}(\mathbf{v}^*_i, \mathbf{v}_{\rho(i)}) \right)$$

where $L^{rank}(\cdot, \cdot)$ refers to metrics such as Recall@K (Manning et al., 2008) while $L^{visual}(\cdot, \cdot)$ measures coverage and precision of the detected visual regions relative to the true ones (Zou et al., 2023) and λ is a relative weighting factor. In our work, we optimize these separately with focus on ranking accuracy. Figure 1 shows the user interface with input P and output $M^*_{\mathcal{A}}(P)$.

3 RxLens Solution Architecture

3.1 Design considerations

Data Privacy. Given the sensitivity of medical data, PII must be robustly redacted from both image and text inputs to third party LLM APIs.

Catalog-based Augmentation. Prescriptions often use medical terms absent in LLM training data. Performing OCR on prescriptions and using the output to retrieve relevant context from medicine catalogs can enhance LLM text interpretation accuracy.

Ensuring Validity of Suggestions. To mitigate medication errors due to LLM hallucination, it is vital to select matching products from the catalog, rather than through direct generation.

Trust and Explainability. To boost customer trust, it is desirable to display relevant visual regions alongside product suggestions.

Low Latency. Given high e-commerce dropout rates, low-latency responses are crucial, even if that entails a slight drop in suggestion quality.

Limited Labeled Data. Prescription digitization spans multiple tasks from medicine extraction to

[1] Catalog refers to a known list of medications.

[2] Mapping ρ can be found based on best match between the visual regions or the medication names across the groups.

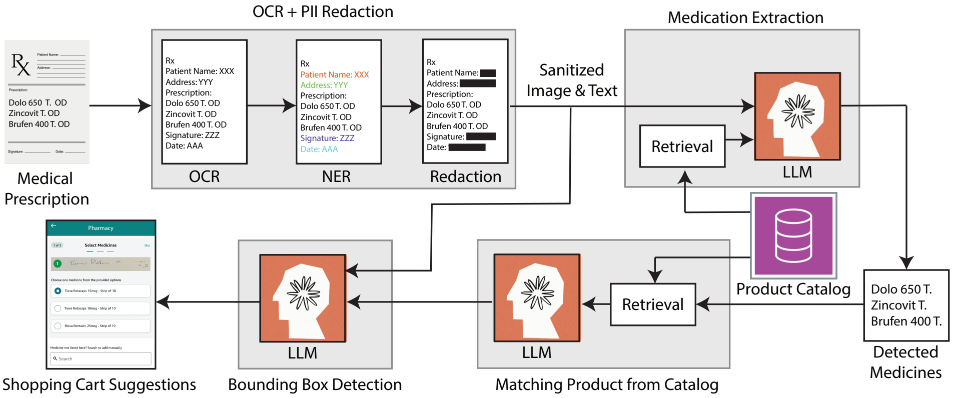


Figure 2: Schematic of the RxLens model pipeline.

catalog validation, each with limited labeled data and significant diversity across market places. Using SOTA LLM APIs with world knowledge, enhanced by contextual retrieval, is likely to be more effective than training custom models.

3.2 Key Processing Stages

Accounting for the above factors, we present our RxLens architecture in Figure 2, which comprises four online processing steps and an offline evaluation step, each optimized via empirical analysis.

PII Redaction and OCR. We first employ secure OCR and named-entity recognition (NER) to extract text from prescriptions, followed by identification and redaction of sensitive PII entities such as names and phone numbers from both text and input images. The sanitized outputs can then be processed via third-party multimodal LLM APIs to improve extraction quality.

Medication Extraction. Using sanitized text and prescription image, we extract medication records with pharmacy-mandated attributes: medicine name, dosage form, and dosage strength. To address vocabulary gaps in generative LLMs during the extraction, we augment the LLM prompt with relevant product titles retrieved from the catalog using the OCR text. To balance extraction accuracy, computational costs, and latency, we optimize input combinations (image, text, catalog context) and prompt design (role, task, format, in-context learning examples) (Chen et al., 2023).

Matching products from Catalog. For each extracted medication, we identify top catalog matches prioritizing ranking accuracy. We explore several retrieval methods, ranging from simple text searches to more complex ones based on weighted attribute similarity. To leverage LLM fuzzy matching capabilities (e.g., matching 0.5g with 500 milligrams), we also consider a three-step retrieval process comprising text search followed by LLM-based ranking, and validation against the catalog.

Table 1: Metrics computed for different tasks within RxLens pipeline and their definitions.

Task (s)	Metric	Definition (average per prescription)
Any	P90 Latency (s)	90th percentile of latency for that task
	Cost (¢)	Cost of AWS Services/LLMs
OCR & Medication Extraction	Medicine-name (M)-Recall	Fraction of ground truth medicines whose attributes (M, M+F, M +F+S) are present in OCR and Medication Extraction output with a "fuzzy" match to permit downstream detection
	Medicine-name+Dosage Form (M+F) Recall	
	Medicine-name+Dosage Form+Strength (M+F+S) Recall	
Matching ASINs from Catalog	Medication-Recall@K	Fraction of ground truth medicines that can be found in final retrieved top K ASIN suggestions with exact match.
PII redaction	Precision & Recall	Precision & recall w.r.t human judgement
Bounding box (BB) Lineage	Coverage & Precision	Fraction of medication groups for which a BB is identified and the where the identified BB overlaps with the ground truth one

Bounding Box Lineage. Finally, we link medication suggestions to visual regions in the prescription, using LLMs to identify the relevant boxes using the OCR output. The smallest rectangle encompassing the relevant boxes is displayed alongside the medication suggestions.

Offline Auto-evaluation. Additionally, we also perform offline auto evaluation of the online processing steps using customer cart preferences as implicit feedback. While the matching against catalog can be directly assessed, for the OCR and medication extraction steps, we use an LLM to estimate the recall of key attributes associated with the user-selected medications within the respective outputs, calibrating it with human judgements.

4 Experimental Setup

We describe our setup for evaluating LLM-based prescription digitization focusing on questions related to solution architecture, component choices, deployment constraints, and auto-evaluation.

4.1 Datasets

To the best of our knowledge, there are no public datasets of unstructured prescription images paired with ground truth digitization. Hence, we use two proprietary e-pharmacy datasets: `Handwritten` and `Printed`, comprising 1469 handwritten and 1001 printed prescriptions respectively. All prescription images undergo PII redaction, customer ID anonymization, and are paired with pharmacist-digitized orders. These prescriptions are sourced from a diverse range of clinics, hospitals, and practitioners from an emerging marketplace, featuring varied formats, abbreviations (e.g., T., Tab. Tablets), layouts (e.g., double column, slanting), image resolutions, and orientations. Since our LLM-based solution(s) and baselines do not involve training, we evaluate each digitization step across the full datasets. To assess offline LLM-based auto-evaluation, we obtain manual judgements of RxLens output on a subset of the data.

4.2 Tasks and Models

As discussed in Section 3, our approach comprises the following tasks: PII redaction, OCR, medication extraction, product matching from the catalog, and bounding box detection, with an overlap in the first two tasks. We explore solutions for each of these tasks using judicious combination of models suited for OCR, NER, LLM, and retrieval limiting our exploration to the representative choices below.

OCR - AWS Textract: An automated OCR service for scanned handwritten and printed documents, supporting English and EU multiple languages.

NER - AWS Comprehend, Comprehend Medical: ML services for natural language understanding, capable of extracting named/PII entities with Comprehend Medical tuned for medical entities.

LLM - Claude V3 and V3.5 Sonnet, Llama 3.1-8B: The most powerful cost-effective generative LLMs hosted on AWS Bedrock featuring long context windows (128K tokens for Llama 3.1 and 200K for the Claude models). Results in Section 5 are based on Claude V3 Sonnet and we provide a comparison across LLMs in Appendix B.

Retrieval - AWS OpenSearch: A fully hosted version of ElasticSearch with advanced real-time retrieval and fuzzy matching over large indexes.

Note that all services used in the RxLens system (AWS Comprehend, Textract, Bedrock) are security-certified for medical applications with guaranteed data encryption at rest and in transit. While AWS Bedrock's terms of service guarantee RxLens data privacy and security, we prefer to redact PII from prescriptions to minimize sensitive data exposure to external LLMs.

4.3 Evaluation Metrics

From a business standpoint, the primary metric of interest is the recall of correct medications within the top-K suggestions (Recall@K), with latency and LLM generation costs being secondary metrics. For proprietary reasons, we skip discussion of the impact of these metrics on operational costs and customer experience. Additionally, we also evaluate various task-level metrics listed in Table 1. At each stage, we evaluate whether the output permits downstream detection of the medicine name, dosage form, and dosage strength of the medications corresponding to the ground truth medicines. We also evaluate the effectiveness of PII redaction, and the accuracy of bounding box mapping for medication suggestions. Lastly, we assess the correlation between LLM-based auto-evaluation and manual judgments.

5 Experimental Results

5.1 Component-wise Design Choices

Below we present evaluation of the design choices associated with the three critical steps of the RxLens digitization pipeline.

OCR. We evaluate two choices: a) `Textract` and b) `OCR-Claude`, which is Claude prompt-tuned for prescription text extraction. Table 2 compares their performance on medication attribute extraction, latency, and compute costs. Surprisingly, `Textract` is not only faster and cheaper but more accurate especially on handwritten prescriptions due to inbuilt correction of image orientation and document image-specific training versus Claude's general-purpose design, making it our preferred choice.

Medication Extraction (Med-Extract). Here, we evaluate three approaches: (a) `Comprehend Medical` (Comp-Med), (b) `Extract-Claude` based on Claude prompt-tuned to extract medication records from the prescription image and OCR output, (c) `Med-Extract-Claude-IR`, which is a `RAG`-variant of Med-Extract-Claude where relevant products from the catalog are identified using an intermediate retrieval (IR) step (matching each line of OCR output with text Jaccard similarity) and included in the prompt as additional context. For approaches (b) and (c), we consider variants with Image-only, Text-only and Image+Text as inputs. Table 2 shows the attribute recall results pointing to clear superiority of Claude-based methods over Comprehend Medical especially on handwritten prescriptions, despite the specialized medical tuning, possibly because of limited coverage of non-US prescriptions in its training data. We observe a sizeable boost due to the inclusion of additional catalog context especially for handwritten prescriptions (+10% medicine name recall) likely due to correction of OCR errors. Including images with the OCR text leads to slightly better extraction but entails extra latency, compute costs and PII redaction effort. Table 4 in Appendix B compares the performance of multiple SOTA LLMs (Claude 3.5 Sonnet, Claude v3 Sonnet, Llama 3.1-8B) on this task.

Matching products against Catalog. We evaluate three approaches: (a) Simple Text Search using Jaccard similarity on medicine names, (b) Attribute Search, which ranks products using a weighted combination of similarities along each attribute (Medicine Name: 2, Dosage Form: 3, Dosage Strength: 2) with weights determined via Bayesian optimization (Perrone et al., 2021), and (c) Reranker-Claude, which combines the output of the first two methods and reranks using Claude. Figure 3 shows the ranking performance in terms of recall@K, pointing to the clear superiority of the re-ranking approach especially at low K due to the LLM's fuzzy matching abilities and *a priori* knowledge on medication attributes.

5.2 Overall Performance vs. SOTA methods

To assess the overall digitization performance of RxLens system, we compare the implementation with optimised choices for each step with two other natural end-to-end baseline systems where the first OCR step is performed using Textract. For the first baseline the latter steps involve Comprehend Medical + Attribute-search for matching, while the second one RAG-Claude is based on conventional retrieval-augmented generation with the first step involving retrieval of relevant products based on the OCR text followed by invocation of Claude, prompt-tuned to perform both medication extraction and the generation of product suggestions while utilising the context. Results in Table 3 point to the dominance of the RxLens approach over the alternatives. Anecdotal results point to the utility of enhancing medication extraction with retrieval augmentation (e.g., Dislar being corrected to Deslor) as well as enhancing ranking with additional LLMs for superior fuzzy matching (e.g., 50 mg matched against 0.05 gram). Superior performance of Rx-Lens relative to RAG-Claude also points to benefits of decomposing a complex task into multiple steps and interleaving retrieval with generation (Khattab et al., 2024).

5.3 Practical System Considerations

For a practical customer-facing system, data privacy, latency, and usability are paramount. Below, we discuss evaluation of our proposed approach for handling these aspects as discussed in Section 3.

PII Redaction. Manual assessment of Comprehend on PII information detection points to a precision and recall of **90.7%** and **82.9%** respectively for printed prescriptions and of **69.4%** and **81.3%** for handwritten prescriptions. Most of the errors can be attributed to personal signature blocks and non-English text, which does not actually pose privacy risk when only the OCR output (and not the sanitised image) is used in the later stages. Further, our choice of PII definitions includes attributes such as gender and age, which by themselves might not be highly sensitive, and are viewed as not PII as per Comprehend contributing to the recall gap.

BB Lineage. We identify the bounding box for each extracted medication using a suitable LLM prompt (Lineage-Claude. Comparing with expert annotations, the coverage for detecting the relevant BBs stands at **75%** and **100%** while the precision of the identified BBs is **87.5%** and **94.1%** for handwritten and printed prescriptions respectively.

Latency Optimization. Since response time is critical in real-time customer-facing flows, we optimised the LLM prompts and inference process by parallelising the retrieval and LLM calls for reranking suggestions for each extracted medicine record, resulting in a 2.5x decrease in overall latency.

5.4 Offline AutoEvaluation using LLMs

Since obtaining fine-grained manual annotations of prescriptions is labour intensive, we explore LLM-based auto evaluation (AutoEval-Claude) of the intermediate stages of RxLens using only the final user-selected product list. We observe correlations ranging from 76% - 88% respectively with human

Table 2: Performance of the different models within the OCR and Extraction phase across the Handwritten and Printed prescription for Medicine-name (M), Medicine-Name + Dosage-Form (M + F) and Medicine-name + Dosage-Form + Dosage-Strength (M + F + S). Note the cost reported is in cents (¢) and Latency is seconds (s).

Phase	Model	Input Type	Handwritten			Printed			Cost (¢)	Latency (s)
			M	M+F	M+F+S	M	M+F	M+F+S		
OCR	Textract	Img	80.6%	65.6%	26.9%	89.0%	85.5%	55.1%	0.15	2.5
	OCR-Claude	Img	54.1%	41.8%	15.9%	76.9%	73.8%	51.5%	0.50	6.5
Med-Extract	Comp-Med.	Txt	14.2%	5.1%	1.4%	62.4%	42.7%	13.2%	0.24	0.9
	No Context	Img	20.9%	16.5%	6.4%	55.5%	50.1%	13.8%	0.42	3.0
		Txt	46.7%	32.6%	14.2%	77.8%	68.3%	32.5%	0.40	2.8
		Img+Txt	47.4%	34.3%	15.2%	79.6%	72.1%	32.8%	0.70	3.5
	IR-Context	Txt	57.3%	38.6%	16.2%	80.6%	71.4%	32.1%	0.44	3.1
		Img+Txt	57.2%	41.2%	18.2%	81.7%	72.9%	30.7%	0.74	3.6

Table 3: Performance comparison of different SoTA approaches (excluding BB lineage step).

Prescription Set	Handwritten		Printed		Overall	
Model	Recall@1	Recall@3	Recall@1	Recall@3	Cost (¢)	Latency (s)
RxLens	38.4%	53.9%	60.2%	75.5%	2.3	12.1
RAG-Claude	25.8%	34.9%	49.9%	64.4%	1.7	3.7
Comprehend Medical	11.1%	13.8%	38.2%	49.9%	0.38	4.4

annotations for Medication Name, Dosage Form, and Strength for the OCR stage and 78% - 100% for the Medicine Extraction stage (see Figure 4). As expected, there is a superior correlation on printed prescriptions relative to handwritten ones. Upon further examination, we find that the divergence primarily arises from fuzzy matching interpretation, with human experts being more lenient than the LLM, suggesting slightly pessimistic yet directionally valid evaluations. Note that our LLM-based auto-evaluation aims to supplement, not replace manual evaluation by enabling robust large-scale monitoring previously limited by manual effort. Expert annotations collected at smaller scale help calibrate and refine the automated system.

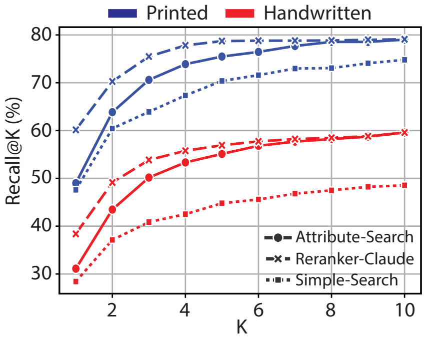


Figure 3: Recall@K vs. K for various retrieval methods across Handwritten and Printed prescriptions.

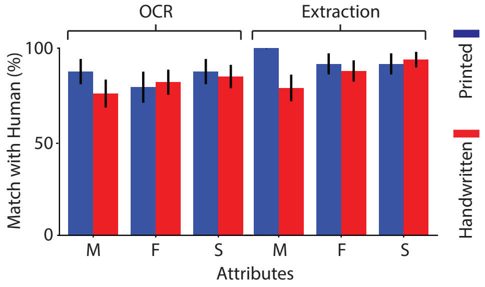


Figure 4: Agreement between AutoEval-Claude and human annotations on the prescription images for Medicine Name (M), Dosage Form (F) and Dosage Strength (S), evaluated across Handwritten and Printed prescriptions. Error bars: Binomial error.

6 Conclusion and Future Work

Our current work presents an LLM-based architecture of a deployed system for digitizing medical prescriptions, assessing various design choices including data privacy and usability.

Summary of key learnings. 1) Specialized models can sometimes outperform foundational models, such as Textract trained on document images outperforming Claude. 2) Retrieval augmentation with relevant context can yield significant performance benefits for specialized domains like pharmacy. 3) Reranking with LLMs improves top ranking results due to their intrinsic world knowledge and ability to perform fuzzy matching over textual attributes. 4) Auto-evaluation using LLMs closely matches human evaluation, enabling scalable monitoring and system optimization. 5) For real-time applications, latency is an important factor, making it crucial to focus on parallelization opportunities.

Future directions. We also plan to explore (a) specialized multimodal models for handwritten content recognition, (b) automated prompt optimization using meta-prompting strategies, (c) assessment of auto-evaluation with more manual annotations. The approach can also be extended to digitizing other documents such as shopping lists.

Limitations

While RxLens has proven fairly effective, it does have some limitations that need to be addressed.

OCR from Handwritten prescriptions. The performance of our current OCR model (AWS Textract) on handwritten prescription data depends on the legibility of the handwriting, with low recall particularly for the strength attribute. To address this, we plan to fine-tune existing handwritten text recognition models on prescription images.

Multilingual support. While all the components of RxLens support multiple languages, our study primarily focused on English-language support, as the medication attributes critical for shopping cart construction are typically written in English even if there is some other non-English content, e.g., medication consumption instructions. For health applications requiring complete prescription digitization, it might be necessary to augment RxLens with multilingual medical vocabularies and perform further evaluation on multilingual support.

Dependence on Catalog Quality and Coverage. Since retrieval augmentation is a critical step in our methodology, the overall performance of RxLens depends heavily on the quality and coverage of the medication catalog used for retrieval. Expanding the catalog to be as exhaustive and standardized as possible is an important area of improvement.

Dependence on LLM choice. Since RxLens involves multiple steps that require invoking a language model, the current prompts used have been optimized for Claude V3 Sonnet. As we explore new LLMs, we will need to automate the process of prompt optimization.

Ethics Statement

Our work aims to expand the adoption of online pharmaceutical services in emerging markets by digitizing medical prescriptions. We are acutely aware of the sensitive nature of prescription data and its potential health impacts, and have taken several steps to ensure the ethical development and deployment of our system as discussed below.

Data Safety. We employ a secure pipeline with appropriate encryption to collect, store, and annotate customer prescriptions. To protect customer privacy and prevent data leakage, we use AWS services (Textract, Comprehend) to detect and redact all personally identifiable information from the prescription text and image before performing LLM-based inference. As we are using a pretrained LLM (Claude), the prescription data is not directly used to train any language model. However, the performance relative to expert digitization is used to optimize system hyperparameters.

System Bias. Pre-trained foundational LLMs are often ill-equipped to handle tasks in specialized domains such as pharmacy due to gaps in their training data. Additionally, these models may have limited exposure to the unique vocabulary and layouts of prescriptions originating from emerging markets, which could hinder their performance if used directly. To mitigate these gaps, our solution design prioritizes retrieval augmentation of LLMs with a region-specific medicine catalog. In future, we plan to continually optimize the prompts and retrieval algorithms based on customer implicit feedback on the suggested medications to further reduce the system biases.

Health Safety. Customer well-being is our top priority. To eliminate the risk of errors that could lead to adverse health impacts, RxLens only presents the top three medication suggestions that meet a certain score threshold, and enables dual review by customers and pharmacists. Highlighting the relevant visual regions in the prescription also helps customers assess the suggestions without undue cognitive load. Our LLM-based auto-evaluation approach paired with suggestion acceptance metrics also also enables the continuous monitoring of system performance and the proactive detection of any issues.

References

Anthropic. 2023. The Claude 3 model family: Opus, Sonnet, Haiku.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tao C, Filannino M, and Uzuner Ö. 2017. Prescription extraction using crfs and word embeddings. Journal of Biomedical information.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Irene Davis and S. FACSM. 2008. Use of real-time feedback to improve dynamic aligment and reduce excessive loading. *Medicine Science in Sports Exercise*, 40(5).

Lovely Joy Fajardo, Niño Joshua Sorillo, Jaycel Garlit, Cia Dennise Tomines, Mideth B. Abisado, Joseph Marvin R. Imperial, Ramon L. Rodriguez, and Bernie S. Fabito. 2019. Doctor's cursive handwriting recognition system using deep learning. In *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–6.

Pavithiran G, Sharan Padmanabhan, Nuvvuru Divya, Aswathy V, Irene Jerusha P, and Chandar B. 2022. Doctors handwritten prescription recognition system in multi language using deep learning. *Preprint*, arXiv:2210.11666.

Mehul Gupta and Kabir Soeny. 2021. Algorithms for rapid digitalization of prescriptions. *Visual Informatics*, 5(3):54–69.

Benedict Guzman, Isabel Metzger, Yindalon Aphinyanaphongs, Himanshu Grover, et al. 2020. Assessment of Amazon Comprehend Medical: Medication information extraction. *arXiv preprint arXiv:2002.00481*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc" Najork. 2020. Representation learning for information extraction from form-like documents. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. Mm1: Methods, analysis insights from multimodal llm pre-training. *Preprint*, arXiv:2403.09611.

Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc*, 17(5):524–527.

Valerio Perrone, Huibin Shen, Aida Zolic, Iaroslav Shcherbatyi, Amr Ahmed, Tanya Bansal, Michele Donini, Fela Winkelmolen, Rodolphe Jenatton, Jean Baptiste Faddoul, Barbara Pogorzelska, Miroslav Miladinovic, Krishnaram Kenthapadi, Matthias Seeger, and Cédric Archambeau. 2021. Amazon sagemaker automatic model tuning: Scalable gagradient-free optimization. *Preprint*, arXiv:2012.08489.

L. Rasmy, Y. Xiang, and Z. Xie. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Nature Digital Medicine*.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training NLP models: A concise overview. *Preprint*, arXiv:2004.08900.

Megha Sharma, Tushar Vatsal, Srujana Merugu, and Aruna Rajan. 2023. Automated digitization of unstructured medical prescriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 794–805.

Ozlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *J Am Med Inform Assoc*, 17(5):514–518.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276.

Appendix A Additional Works

Prescription digitization has attracted increasing attention as a vital prerequisite for digital transformation of healthcare services. Most earlier methods (Guzman et al., 2020; Uzuner et al., 2010; Patrick and Li, 2010), focus on entity recognition assuming input is unstructured text and evaluate on printed clinical documents from US. Recent techniques (Sharma et al., 2023; G et al., 2022; Rasmy et al., 2021; C et al., 2017) address the task of digitizing images of paper prescriptions using Convolutional Neural Networks (CNNs) or off-the-shelf tools such as Textract for OCR. This step is followed by further analysis of the OCR output (text and positional information) using sequence fine-tuned models such as Recurrent Neural Networks (RNNs), LSTMs and more recently Transformer models such as BERT and LayoutLM combined with Conditional Random Fields (CRFs) to detect the medication attributes such as medication names, and dosages, along with their associations. These techniques based on custom models, however, require substantial manual annotations.

Document AI primarily deals with understanding visually rich documents (VRDs) by combining compute vision techniques with layout and text understanding. While these techniques (Barrow et al., 2020; Katti et al., 2018; Majumder et al., 2020; Cui et al., 2021) based on graph neural networks and layout-enhanced Transformer models are effective in extracting structured data from well-formatted printed documents with tables such as invoices, these perform poorly on handwritten documents and heterogeneous layouts. Increasingly, these techniques are being replaced by the more versatile multimodal LLM solutions.

Multimodal Generative LLMs such as GPT-4, Claude (Anthropic, 2023) that can process both textual and visual data have emerged as powerful automation and analysis tools. In principle, these models can be directly prompted to digitise a prescription image and convert it to into a list of canonicalised products in a single invocation. However, in practice, the resulting digitization quality is fairly low since these foundational models have scant exposure to medical vocabulary and handwritten prescription images. Currently, even the OCR performance of these models on medical documents lags behind simpler models though that is likely to change over time. Solution strategies typically involve decomposing complex tasks and combining MLLM invocation with additional preprocessing, retrieval, and post processing steps (Khattab et al., 2024). In our current work, we employ Claude V3 Sonnet (Anthropic, 2023) multimodal system to digitize both printed and handwritten medical prescription utilising a similar multi-step strategy including retrieval from medical knowledge base to allow the LLM to reason about the context of medical terminology and abbreviations and improve extraction accuracy.

Appendix B Comparison across LLMs

Table 4 compares the performance of different large language models (LLMs) in extracting medical information, specifically medicine names (M), medicine names with dosage forms (M+F), and medicine names with both dosage forms and dosage strengths (M+F+S), from both handwritten and printed prescriptions. The models evaluated are Claude Sonnet v3, Claude Sonnet v3.5, and Llama 3.1 8b, with performance metrics shown for both handwritten and printed inputs.

Overall, the Claude Sonnet models demonstrate more robust performance across both handwritten and printed prescriptions, with slight improvements observed in the transition from v3 to v3.5. In contrast, Llama 3.1 8b tends to underperform in comparison, especially when the extraction task includes both dosage form and dosage strength.

Table 4: Comparison of LLMs in the Extraction phase for retrieving context from catalog and text-only inputs across Handwritten and Printed prescriptions for M, M+F, M+F+S. (M = Medicine-name, F = Dosage-Form, S = Dosage-Strength)

Model	Handwritten			Printed		
	M	M+F	M+F+S	M	M+F	M+F+S
Claude Sonnet v3	57.3	38.6	16.2	80.6	71.4	32.1
Claude Sonnet v3.5	58.4	38.5	16.5	81.5	71.5	32.2
Llama 3.1 8b	55.1	40.6	17.1	74	64.8	23.8

Appendix C API Costs

Table 4 provides additional details on the average cost of invoking various AWS services and Claude V3 Sonnet for different tasks.

Task	**API**	**Char.**	**Img Size**	**Input Tokens**	**Output Tokens**	**Cost (¢)**
OCR	Claude Sonnet	-	0.74	126	116	0.508
Extract-Img	Claude Sonnet	-	0.74	240	38	0.425
Extract-Txt	Claude Sonnet	-	0	1182	31	0.401
Extract-Img+Txt	Claude Sonnet	-	0.74	1201	33	0.706
ExtractIR-Img	Claude Sonnet	-	0.74	304	34	0.438
ExtractIR-Img+Text	Claude Sonnet	-	0.74	1318	33	0.741
Reranker	Claude Sonnet	-	0	1216	664	1.361
RAG	Claude Sonnet	-	0	1523	681	1.478
OCR	Textract	-	-	-	-	0.15
NER	Comprehend	946	-	-	-	0.095
NER	Comprehend Medical	946	-	-	-	0.237

Table 5: This Table provides additional details on the average cost of invoking various AWS services and Claude V3 Sonnet for different tasks. The cost (in ¢) was computed based on the following pricing policy. **Claude V3 Sonnet**: $3 per Million input tokens, $15 per Million output tokens, $4 per 1000 1MP images. **AWS Textract**: $1.5 per 1000 pages. **AWS Comprehend**: $1 per Million characters. **AWS Comprehend Medical - RxNorm**: $2.5 per Million characters

Appendix D Prompt Templates

Algorithm 1 Medical Prescription Extraction Prompt Template

```
Role: Define the role description for the task (e.g., Medical Assistant, Prescription Interpreter, etc.)
Task: Define the task description including the rules, relevant domain information, and the expected
  input-output format.
Input:
    • OCR Output: Text captured from the scanned prescription.
    • Prescription Image: The scanned prescription.
    • Medicine List: List of possible relevant medicine names retrieved from the catalog.
Output: Expected output format: A structured list with the name of the medicine, its dosage form,
  and its strength.
In-Context Learning Examples:
    • Input: OCR output + image of a medical prescription + list of possible medicine names.
    • Output: A formatted list of medicines with the following fields:
        – Name of the medicine.
        – Dosage form (e.g., tablet, suspension, etc.).
        – Strength (e.g., 500mg, 1g, etc.).
Steps:
    1. Extract relevant data from OCR output.
    2. Cross-reference extracted data with medicine catalog.
    3. Format the output to list medicines, their dosage form, and strength.
    4. Ensure all fields are clearly separated and properly formatted.
Output Format: List of medicines with columns for:
    • Name
    • Dosage Form
    • Strength
```