A Comparative Evaluation of Text Detection and Recognition Models for Cross-hospital Generalization of Thai Pharmaceutical Label OCR under Limited Data Conditions

Wongpanya S. Nuankaew [1], Pathapol Jomsawan [2], Kuljira S. Nuankaew [3], Kaewpanya S. Nuankaew [3], and Pratya Nuankaew [1,*]

[1] School of Information and Communication Technology, University of Phayao, Phayao, Thailand
[2] Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
[3] Phayao Pittayakhom School, Phayao, Thailand
Email: wongpanya.nu@up.ac.th (W.S.N.); pathapon_j@cmu.ac.th (P.J.); nuankaew.kj@gmail.com (Ku.S.N.); nuankaew.kp@gmail.com (Ka.S.N.); pratya.nu@up.ac.th (P.N.)
*Corresponding author

***Abstract*—Digitizing pharmaceutical labels using Optical Character Recognition (OCR) can improve medication safety and hospital workflow efficiency, but cross-hospital deployment remains challenging due to variations in label templates, printing formats, and limited labeled medical data. This study evaluates text detection and recognition models for Thai pharmaceutical label OCR under cross-hospital conditions using a three-tier evaluation protocol: source-hospital data, mixed-hospital data, and unseen hospitals. The results show clear architectural differences in cross-hospital robustness. The PaddleOCR recognition model achieves the best cross-hospital performance with 80.47% accuracy on unseen hospitals, while Transformer-based TrOCR variants degrade substantially under domain shift, achieving only 30–52% accuracy. Text detection shows relatively stable transfer across hospitals with a moderate drop of −10.68%, whereas recognition performance varies notably across architectures and training strategies. Conservative fine-tuning improves recognition accuracy by 14.85% over the base model while preserving cross-hospital generalization. These findings suggest that pre-training diversity and architectural inductive biases are more important than model scale for OCR deployment in limited-data medical settings.**

***Keywords*—cross-hospital generalization, medical document processing, optical character recognition, pre-training quality, Thai script**

I. Introduction

The digitization of pharmaceutical label information using Optical Character Recognition (OCR) is increasingly important as hospitals adopt digital healthcare systems. Medication instructions printed on drug bags often need to be manually transcribed into electronic health records or hospital information systems, a process that is time-consuming and prone to transcription errors. Automated OCR can therefore improve medication safety, reduce manual workload, and support scalable medical data processing. Recent studies have explored OCR-based transcription of pharmaceutical labels to assist clinical information management and medication documentation [1].

Applying OCR to Thai pharmaceutical labels presents challenges distinct from general OCR tasks. The Thai writing system contains complex character compositions, stacked diacritics, tonal markers, and context-dependent glyph placement, which complicate segmentation and recognition. In addition, pharmaceutical labels vary considerably across hospitals due to differences in printing systems, templates, fonts, and layouts. As a result, labels with similar medical content may appear visually different across institutions, causing domain shift when OCR systems are deployed in new hospitals. In practice, labeled datasets are also limited because of privacy regulations and annotation costs, further restricting model training and cross-domain generalization.

Recently OCR architecture has achieved strong performance on benchmark datasets. Lightweight document OCR systems such as PaddleOCR combine convolutional feature extraction with efficient recognition modules designed for practical deployment [2, 3]. Advances in scene text recognition have also introduced architectures such as Scene Text Recognition with Visual Transformer (SVTR) that integrate convolutional and Transformer components for improved representation learning [4]. In parallel, Transformer-based approaches such as TrOCR employ Vision Transformer encoders with autoregressive decoders to leverage large-scale

Manuscript received January 7, 2026; revised February 20, 2026; accepted April 10, 2026; published August 26, 2026.

pre-training for text recognition [5–7]. Despite these advances, most evaluations rely on benchmark datasets where training and testing data share similar distributions. This setting does not reflect real-world medical deployment where models trained in one hospital must operate on pharmaceutical labels from different hospitals.

Several research gaps remain. First, systematic evaluation of OCR architectures under cross-hospital domain shift is limited, particularly for complex scripts such as Thai where document appearance varies across institutions. Second, the robustness of OCR pipeline components, specifically text detection and text recognition, across domains has not been thoroughly examined. Modern OCR systems commonly use cascade pipelines in which detection identifies text regions before recognition converts cropped regions into character sequences [2, 6, 8]. Third, the effects of architectural design and training strategies under limited-data conditions typical of medical environments remain insufficiently studied. Although OCR research has advanced rapidly, most studies emphasize benchmark performance rather than cross-institution deployment. To the best of current knowledge, systematic investigation of cross-hospital OCR generalization for Thai pharmaceutical labels under realistic limited-data conditions remains scarce.

To address these challenges, this study conducts an empirical evaluation of OCR models for Thai pharmaceutical label recognition under cross-hospital deployment conditions. A three-level evaluation protocol is applied, including source-hospital validation, mixed-hospital testing, and unseen hospital evaluation, enabling controlled assessment under increasing domain variation. The main novelty of this work lies in the structured cross-hospital evaluation framework, which enables systematic analysis of OCR generalization and reveals how different OCR architectures perform under real medical deployment conditions. Within this setting, hybrid Convolutional neural network (CNN)–Transformer recognition models implemented in PaddleOCR [2–4] are compared with Transformer-based architectures represented by TrOCR [5]. The main contributions of this study are summarized as follows:

- A structured three-level evaluation framework is proposed to assess OCR generalization across hospitals under increasing domain shift, using source-hospital validation, mixed-hospital testing, and fully unseen hospital data.
- A comparative analysis of OCR pipeline components examines the robustness of text detection and recognition in cross-hospital settings, showing that performance degradation is mainly caused by recognition errors rather than detection failures.
- An empirical study evaluates how different OCR architectures and fine-tuning strategies perform in limited-data medical settings, demonstrating that lightweight hybrid models such as PaddleOCR generalize better than large Transformer-based models.
- An analysis of accuracy and efficiency trade-offs shows that PaddleOCR Mobile provides the best balance between cross-hospital recognition accuracy and inference latency, supporting its suitability for real-world hospital deployment.

Finally, the remainder of this paper is organized as follows. Section II reviews related work on OCR architecture and medical document recognition. Section III presents the evaluation framework and experimental setup. Section IV reports the experimental results and analysis, followed by discussion and conclusions in the final sections.

II. Literature Review

Recent advances in deep learning have significantly improved OCR systems. This section reviews representative studies on OCR architectures, detection and recognition pipelines, and document analysis applications, with emphasis on challenges related to domain variation and medical document processing.

A. Deep Learning Architectures for OCR

Advances in Optical Character Recognition (OCR) have been driven by deep learning architectures that combine convolutional neural networks with Transformer-based representations. The PaddleOCR framework provides a modular OCR pipeline integrating convolutional feature extraction with sequence-based recognition modules for efficient document processing [2, 3]. In this pipeline, detection networks locate candidate text regions, while recognition modules convert cropped regions into character sequences.

Subsequent research has explored improved recognition architectures to enhance feature representation and sequence modeling. The Scene Text Recognition with Visual Transformer (SVTR) architecture introduces Transformer-based feature modeling into the recognition stage to improve representation learning for complex text patterns. In parallel, Transformer-based OCR models such as TrOCR use a Vision Transformer encoder with an autoregressive decoder for end-to-end text recognition [5]. These architectures leverage self-attention and large-scale pretraining to improve performance across diverse text datasets [6, 7, 9].

Although these approaches achieve strong results on benchmark datasets, evaluations typically focus on standard scene-text benchmarks or large curated datasets. Consequently, the behavior of OCR architectures under domain shift and limited-data conditions remains insufficiently studied.

B. Text Detection and Recognition Pipelines

Most modern OCR systems adopt a cascade pipeline consisting of text detection followed by text recognition. Detection models localize text regions within images, while recognition modules convert detected regions into text sequences. Detection approaches such as the EAST detector and differentiable binarization methods have shown strong performance in scene-text localization tasks [8, 10].

Recognition modules typically use sequence modeling techniques such as Connectionist Temporal Classification (CTC) or attention-based decoders to translate visual features into character sequences [10, 11]. Advances in deep neural networks have further improved recognition accuracy through enhanced feature extraction and sequence modeling [12, 13].

Despite these advances, most studies evaluate detection and recognition components separately or focus on improving individual modules. Limited work analyzes how these components interact under domain variation, particularly in document recognition scenarios where detection and recognition errors may propagate within the OCR pipeline.

C. Medical Document OCR Applications

OCR has been applied to various document analysis tasks, including the digitization of medical records and pharmaceutical labels. Prior studies have examined OCR-based transcription of pharmaceutical labels to support hospital information systems and medication management workflows [1]. In these applications, OCR systems must accurately extract medication names, dosage instructions, and other clinically relevant information from printed labels.

However, medical document OCR presents challenges less common in general OCR benchmarks. Pharmaceutical labels often vary in layout, printing quality, and domain-specific terminology, which can affect recognition accuracy. Several studies have explored machine learning approaches to improve document recognition and information extraction in healthcare applications [14, 15]. Nevertheless, most evaluations rely on single-domain datasets, limiting understanding of OCR performance across multiple institutions.

D. Domain Shift and Robust OCR Deployment

Domain shift remains a major challenge for real-world OCR deployment. Differences in document templates, font styles, image quality, and printing conditions can cause performance degradation when OCR models are applied to new environments. This issue is particularly relevant in healthcare, where documents produced by different hospitals often follow different formatting standards.

Recent studies have explored methods to improve the robustness and generalization of machine learning systems under distribution shift [16–18]. However, systematic evaluation of OCR architectures in cross-hospital deployment scenarios is still limited. Little attention has been given to how different OCR architectures and pipeline components perform under domain shift when labeled training data are scarce.

To position this study within existing literature, representative OCR approaches are summarized in Table I. The comparison highlights differences in architectures, application domains, and evaluation settings, and indicates whether domain variation and detection–recognition interactions are considered.

TABLE I. COMPARISON OF REPRESENTATIVE OCR STUDIES AND THE PROPOSED STUDY

Study	Architecture	Domain	Evaluation Type	Domain Shift	Pipeline Analysis	Focus
Pharmaceutical Label OCR [1]	Commercial OCR engines	Medical labels	Single-hospital deployment	×	×	Feasibility of OCR transcription for medication labels
PaddleOCR Framework [2, 3]	CNN + Transformer hybrid	Document OCR	Benchmark evaluation	×	Partial	Efficient OCR pipeline for document processing
SVTR [4]	Transformer-based recognition	Scene text recognition	Benchmark evaluation	×	Recognition	Improved representation learning for scene text recognition
TrOCR [5]	Vision Transformer encoder–decoder	Scene text recognition	Benchmark evaluation	×	Recognition	State-of-the-art recognition performance on OCR benchmarks
EAST / DB Detection [6, 8]	CNN-based detection networks	Scene text detection	Benchmark evaluation	×	Detection	Accurate text localization
Deep Sequence Recognition [9–14]	CNN-RNN / CTC recognition	Scene text recognition	Benchmark evaluation	×	Recognition	Sequence modeling approaches for OCR recognition
Document OCR Applications [14, 15, 19]	Deep learning OCR pipelines	Document processing	Domain-specific datasets	Limited	Partial	OCR-based document information extraction
Robust ML / Domain Shift Studies [16–18]	General ML robustness methods	Various domains	Domain-shift benchmarks	✓	×	Improving model robustness under distribution shift
Proposed Study	PaddleOCR vs TrOCR	Thai pharmaceutical labels	Cross-hospital evaluation	✓	✓	Systematic analysis of OCR robustness under domain shift

As shown in Table I, most existing OCR studies mainly evaluate architectural performance on benchmark datasets or single-domain document collections. Although these studies report strong recognition accuracy, limited attention has been given to OCR robustness in real-world deployment where document appearance varies across institutions. Cross-hospital domain variation and the interaction between text detection and recognition components remain insufficiently studied. These limitations motivate the present study, which systematically examines OCR performance for Thai pharmaceutical labels under cross-hospital deployment using a multi-level evaluation framework.

Beyond recognition accuracy, deploying OCR systems in healthcare also raises concerns about data security and patient privacy, as medical OCR pipelines process sensitive information such as prescription details, patient identifiers, and dosage instructions subject to data protection regulations. In cross-institution deployments, secure transmission and storage of recognized data are therefore essential. Recent research has explored privacy-preserving mechanisms for protecting medical data in interconnected healthcare systems; for example,

Dhinakaran *et al.* [20] analyze quantum-based privacy-preserving techniques for secure Internet of Medical Things (IoMT) environments, highlighting the role of quantum cryptography and secure communication protocols. Although this study focuses on OCR accuracy and cross-domain generalization for pharmaceutical label recognition, such privacy-preserving frameworks remain important complementary components for real-world OCR-based medical information systems [20].

Overall, prior research has advanced OCR architectures and improved recognition accuracy on benchmark datasets. However, most studies emphasize benchmark-based evaluation and often examine detection or recognition modules independently. Systematic evaluation of OCR robustness under cross-institution deployment remains limited, especially for medical documents where domain variation and limited training data are common. These limitations motivate the present study, which evaluates OCR performance for Thai pharmaceutical labels under cross-hospital deployment using a multi-level evaluation framework.

III. Materials and Methods

A cascade OCR pipeline is implemented in which pharmaceutical label images are first processed by a text detection module to localize text regions, followed by a text recognition module that converts the detected regions into character sequences. The detection stage uses text-localization object detection models, while the recognition stage employs sequence-to-sequence models designed for Thai script, as illustrated in Fig. 1, while the evaluation and training procedures are formally described in Algorithms 1 and 2.

Pharmaceutical label images are collected from a primary source hospital (Hospital A) and multiple target hospitals to construct a cross-healthcare facility dataset. The dataset is divided into training, validation, mixed-domain test, and unseen sets under a three-level evaluation protocol. Images are processed through the cascade OCR pipeline, and performance is evaluated using detection metrics and recognition metrics to assess cross-domain generalization, as shown in Fig. 2.

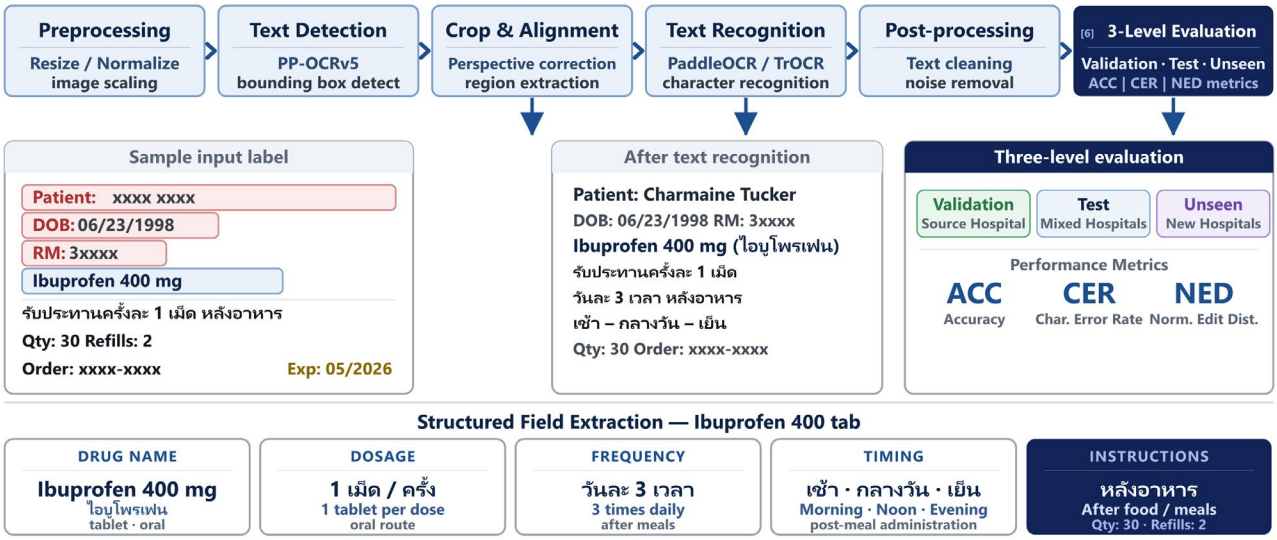


Fig. 1. OCR pipeline for pharmaceutical label processing.

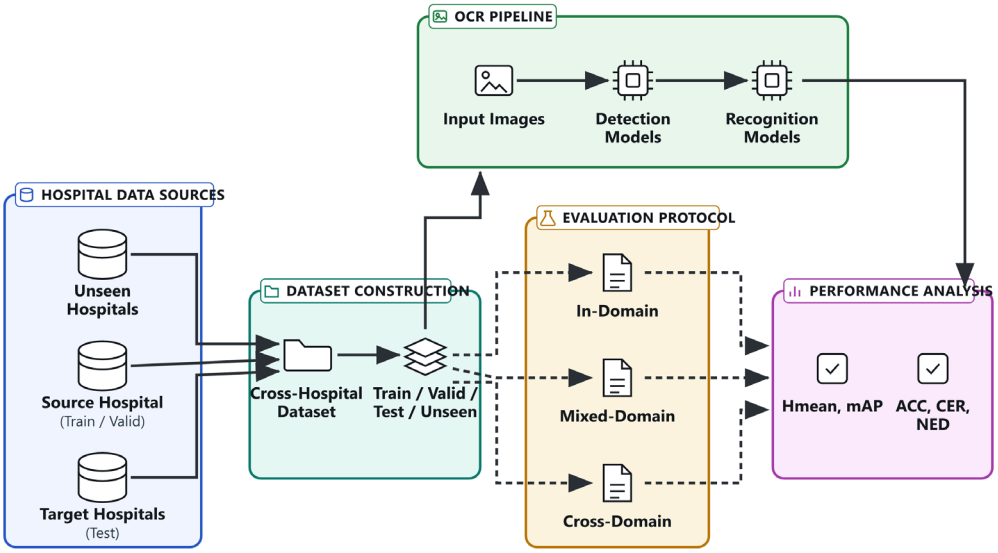


Fig. 2. Cross-hospital experimental design and OCR evaluation framework.

This section describes the datasets, OCR models, evaluation protocol, and training procedures used in this study. The methodology is designed to analyze the generalization behavior of OCR systems when deployed across multiple hospitals with different document layouts and printing conditions.

A. Dataset Collection

Three separate datasets were created to assess various pipeline components across different healthcare facility scenarios. The dataset splits were designed to mimic real-world deployment challenges, incorporating three levels of domain difficulty.

Text Detection Dataset: The dataset contains 403 pharmaceutical label images annotated with bounding boxes. Following the three-level evaluation protocol, 322 images (80%) from Hospital A are used for training and 81 images (20%) for validation. The test set (50 images) contains a mixed-domain composition, with approximately half from Hospital A and half from other healthcare facilities. The unseen set (50 images) consists entirely of images from different hospitals not included in training. The dataset details are summarized in Table II.

Text Recognition Dataset: The dataset contains 3,294 cropped pharmaceutical label images capturing information such as drug names, dosages, and dates, collected at the University of Phayao Hospital. The training set (2,635 images, 80%) and validation set (659 images, 20%) are sourced from Hospital A, where annotation resources are available. The test set (312 images) includes a mixed composition: approximately half from Hospital A and half from other hospitals (e.g., Hospitals B and C) in Chiang Rai and Chiang Mai, simulating deployment in both familiar and new environments. The unseen set (300 images) is collected from different hospitals (e.g., Hospitals D, E, and F) across Chiang Rai, Chiang Mai, and Phayao, containing diverse templates, fonts, layouts, and capturing conditions to evaluate cross-domain generalization. The dataset distribution is summarized in Table II.

TABLE II. DATASET STATISTICS AND HOSPITAL COMPOSITION

Dataset Type	Train (80%)	Valid (20%)	Test	Unseen
Text Recognition	2,635	659	312	300
Text Detection	322	81	50	50

A three-level evaluation framework assesses model generalization under cross-hospital conditions. Images from one hospital serve as the source domain for training and validation, while images from other hospitals are reserved for evaluation under domain shift. The dataset is divided into three subsets: the source-hospital set (D_a) for training and validation, the mixed-hospital set (D_{mix}) with approximately equal samples from the source and other hospitals, and the unseen-hospital set (D_{unseen}) containing images exclusively from hospitals not seen during training.

These subsets represent increasing difficulty: the validation set (easy) evaluates in-domain performance, the mixed-hospital test set (intermediate) measures partial domain shift, and the unseen set (hard) tests full cross-domain robustness. Examples of label images and cropped text regions are shown in Fig. 3, while in-domain and cross-domain samples are shown in Fig. 4.

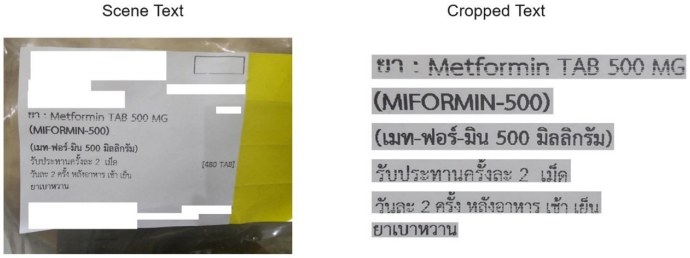


Fig. 3. Sample pharmaceutical labels and cropped text regions.

Fig. 4. In-domain and cross-domain pharmaceutical label examples.

Dataset limitations and privacy compliance: The dataset used in this study is relatively small compared with large-scale OCR benchmarks, mainly due to privacy regulations and institutional approval requirements that restrict access to medical images containing sensitive information.

Despite its size, the dataset includes samples from multiple hospitals with diverse printing templates, layouts, and font styles, enabling evaluation under realistic cross-hospital variation. Accordingly, the experiments focus on analyzing model robustness to domain shift in limited-data medical settings, reflecting practical deployment conditions in real hospital systems.

This study follows ethical guidelines for medical data research with patient consent. All pharmaceutical label images were fully anonymized to comply with Thailand's personal data protection act; patient identifiers were removed using automated redaction with manual verification, and dates were generalized to year–month resolution to reduce re-identification risk while preserving OCR-relevant structure. The dataset is stored on encrypted, access-controlled servers and is not publicly released.

To improve clarity and reproducibility, the cross-hospital OCR evaluation procedure is formalized in Algorithm 1, defining three evaluation settings with increasing levels of distribution shift between training and evaluation data. The protocol is applied separately to text detection and text recognition. For zero-shot models, the fine-tuning steps are skipped.

Proposition: Expected Cross-Hospital Performance Ordering.

Let P_{val}, P_{test}, and P_{unseen} denote OCR performance on the validation, mixed-hospital, and unseen-hospital datasets, respectively. Under a cross-hospital distribution shift, performance typically follows [21].

$$P_{val} \geq P_{test} \geq P_{unseen} \tag{1}$$

where P_{val} denotes in-domain evaluation, P_{test} corresponds to partial domain shift, and P_{unseen} represents fully out-of-domain evaluation.

Proof Sketch: Training and validation data originate from the same source-hospital distribution D_A, leading to the highest expected performance. The mixed-hospital set introduces moderate domain shift from additional hospitals, while the unseen-hospital set exhibits the largest distribution divergence, resulting in the lowest performance. This pattern is consistent with findings in domain adaptation and dataset shift research [21–23].

Algorithm 1 describes the proposed three-level cross-hospital evaluation framework. The source-hospital dataset is first split into training and validation subsets, after which the OCR model is fine-tuned with early stopping to select the best-performing configuration. The resulting model is then evaluated under three progressively challenging scenarios: in-domain validation (D_{val}), mixed-domain testing (D_{mix}), and fully unseen hospitals (D_{unseen}).

Algorithm 1. Three-Level Cross-Hospital OCR Evaluation

Input

D_A: labeled image set from source hospital
D_{mix}: mixed-domain dataset
D_{unseen}: dataset from previously unseen hospitals
M: OCR model
η: learning rate
E_{max}: maximum training epochs

Output

$\{P_{val}, P_{test}, P_{unseen}, \Delta_{abs}, \Delta_{rel}, val_{gap}\}$

1. Partition D_A into D_{train} and D_{val}
2. Fine-tune M on D_{train} with learning rate η using early stopping on D_{val}
3. Record best model weights M^* at the epoch with highest validation accuracy
4. $P_{val} \leftarrow Evaluate(M^*, D_{val})$
5. $P_{test} \leftarrow Evaluate(M^*, D_{mix})$
6. $P_{unseen} \leftarrow Evaluate(M^*, D_{unseen})$
7. $\Delta_{abs} \leftarrow P_{test} \leftarrow P_{unseen}$
8. $\Delta_{rel} \leftarrow \frac{P_{test} - P_{unseen}}{P_{test}} \times 100\%$
9. $val_{gap} \leftarrow P_{val} \leftarrow P_{unseen}$
10. Return $\{P_{val}, P_{test}, P_{unseen}, \Delta_{abs}, \Delta_{rel}, val_{gap}\}$

To assess generalization under domain shift, both absolute (Δ_{abs}) and relative (Δ_{rel}) performance degradation is computed, together with the validation gap (val_{gap}). This evaluation protocol supports systematic analysis of model robustness and cross-domain transferability in real-world deployment settings.

B. Text Detection Models

Within the PaddleOCR framework, text detection functions as a standalone module for localizing text regions in pharmaceutical label images. This study employs PP-OCRv5 Mobile detection models to examine the effect of domain-specific fine-tuning on cross-domain transferability in real-world deployment. The method is based on the Differentiable Binarization (DB) framework, which jointly estimates probability and threshold maps and combines them via approximate binarization to produce accurate text region segmentations.

PaddleOCR's modular, document-aware design enables a decoupled OCR pipeline, supporting controlled and interpretable evaluation of text recognition models.

1) Evolution of PaddleOCR

Since its release in 2020, PaddleOCR has evolved alongside advances in OCR and document understanding. Early versions (1.x–2.x) were integrated end-to-end systems emphasizing accuracy, inference speed, and deployment simplicity. The PP-OCR series (v1–v4) adopted a tightly coupled OCR-centric design, effective for text transcription but limited in scalability, flexibility, and extensibility as OCR outputs became increasingly used in downstream tasks. As OCR results became essential for document parsing, key information extraction, and retrieval-augmented generation, the need for structured and reusable representations increased. PaddleOCR 3.0 addresses these challenges by introducing a modular, pipeline-based document intelligence

architecture, where OCR is one component within a broader workflow.

The system comprises three independently deployable modules: PP-OCRv5 for text recognition, PP-StructureV3 for document parsing, and PP-ChatOCRv4 for LLM-assisted information extraction. This design enables flexible configuration and task-specific optimization across diverse applications [19].

2) PP-OCRv5 model

PP-OCRv5 follows the modular design of PaddleOCR 3.0, operating as an independent and reusable text detection component within a layered OCR system. It adopts PP-HGNetV2 as the backbone, providing enhanced feature representation while maintaining a lightweight structure suitable for real-world deployment. To improve generalization without increasing inference cost, the model incorporates knowledge distillation by aligning its features with a stronger teacher model, GOT-OCR2.0. It also retains established strategies, including Parallel Fusion Head (PFHead) and Dynamic Scale-aware Refinement (DSR), to ensure robust multi-scale text detection.

Robustness is further improved through extensive data augmentation and hard-case mining, particularly at the text-line level and across scripts, enabling resilience to variations in font, orientation, blur, and geometric distortion. These design choices make PP-OCRv5 well-suited for practical, resource-constrained environments. In this study, PP-OCRv5 is fixed across all experiments to minimize detection variability, ensuring consistent text localization and enabling fair comparison of recognition models [19].

A decoupled OCR pipeline is adopted, where PP-OCRv5 performs detection and multiple recognition models are evaluated on identical text regions. This removes detection-induced variance and allows performance differences to be attributed solely to recognition architectures and training strategies, consistent with the modular design of PaddleOCR 3.0.

C. Text Recognition Models

Text recognition transforms cropped text regions into corresponding character sequences. To assess cross-domain generalization under data-scarce conditions, the recognition stage is evaluated independently of detection by comparing two distinct model paradigms.

The first paradigm consists of lightweight CNN–CTC-based models implemented via PaddleOCR Mobile, assessed in both baseline and fine-tuned configurations. These models adopt a hybrid CNN–Transformer architecture designed for computational efficiency. Specifically, PPLCNetV3 serves as the backbone for lightweight feature extraction, while SVTR (Scene Text Recognition with Vision Transformer) captures long-range dependencies through self-attention mechanisms. Sequence prediction is performed using Connectionist Temporal Classification (CTC), which enables alignment-free training. This architecture provides a favorable trade-off between accuracy and efficiency, with a compact model size of approximately 6 MB, making it suitable for deployment scenarios [2, 3].

The second paradigm comprises Transformer-based encoder–decoder architectures derived from TrOCR, including openthaigpt/thai-trocr and kkatiz/thai-trocr-thaigov-v2, also evaluated under both baseline and fine-tuned settings. Input images are resized to 384×384 pixels and partitioned into 16×16 patches. The encoder employs a Vision Transformer (ViT) to extract visual representations via multi-head self-attention, while the decoder generates text sequences autoregressively using masked self-attention in conjunction with cross-attention [5].

The OpenThaiGPT/Thai-Trocr model is initialized from OpenThaiGPT, enabling general-purpose Thai language modeling. In contrast, thai-trocr-thaigov-v2 is pre-trained on 250,000 synthetic images of Thai government documents, providing domain-specific prior knowledge. Both models rely exclusively on self-attention mechanisms and utilize beam search for decoding, in contrast to the greedy CTC decoding employed by PaddleOCR [5].

D. Fine-Tuning and Implications for Modular OCR Design

Recognition models are evaluated under both baseline and fine-tuned configurations while employing a fixed detection backbone, allowing the effects of domain adaptation to be isolated. The impact of fine-tuning varies across architectures: CNN–CTC-based models primarily enhance visual robustness, whereas Transformer-based models additionally gain from language-level adaptation, which is particularly critical for Thai and domain-specific text. These findings underscore the benefits of modular OCR architectures, such as PaddleOCR 3.0, where decoupled pipelines facilitate flexible adaptation and enable consistent cross-domain evaluation.

A conservative fine-tuning strategy is used to adapt pre-trained OCR models to the pharmaceutical label dataset. The training procedure is outlined in Algorithm 2. The OCR pipeline contains two stages: text detection and text recognition.

1) Detection

The DB-based detector operates on an input image of size $H \times W$, where H and W denote the image height and width, respectively, and Crepresents the number of feature channels. Its computational complexity is O(HWC), which increases linearly with the image resolution [24].

2) Recognition

Cropped text sequences of length L are decoded using CTC-based or Transformer-based models [9].

- CTC-based models: $O(L|\Sigma|)$.

where $|\Sigma|$ denotes the size of the vocabulary (character set).

- Transformer-based models: $O(L^2 d)$, where d corresponds to the model's hidden dimension.

Given that pharmaceutical label text is relatively short (≈ 15–40 characters), the quadratic attention cost remains manageable.

3) Evaluation and training complexity

Evaluation Metrics such as Character Error Rate (CER) and Normalized Edit Distance (NED) compute the Levenshtein distance between predicted and ground-truth sequences. Using dynamic programming, the cost per sample is $O\left(\left|y\right|\left|\hat{y}\right|\right)$ [25].

The overall fine-tuning cost in Algorithm 2 is approximated as $O(E\times N\times C_{model})$, where E is the number of training epochs, N is the number of training samples, and C_{model} denotes the architecture-dependent cost of forward and backward gradient computation. This follows the standard view of stochastic gradient optimization, where total training cost scales with the number of iterations and the per-iteration gradient computation cost [26].

Algorithm 2. Conservative Fine-Tuning under Limited-Data Cross-Hospital OCR

Input

M_0: pre-trained OCR model

D_{train}, D_{val}: source-hospital splits, where $\left|D_{train}\right| = N$

$\eta \in \left\{10^{-5}, 10^{-4}\right\}$: learning rate

E_{max}: maximum training epochs

L: task loss function

patience: early stopping threshold

Output

Fine-tuned model M^*

Initialize $M \leftarrow M_0$
Set *best_val_acc* ← 0, *no_improve* ← 0
for *epoch* = 1 to E_{max} **do**
  for each mini-batch B in D_{train} **do**
    $\hat{y} \leftarrow M(B)$
    loss ← $L(\hat{y}, y)$
    Update M by gradient descent with learning rate η
  end for
  val_acc ← Evaluate(M, D_{val}).*ACC*
  if *val_acc* > *best_val_acc* then
    best_val_acc ← *val_acc*
    M^* ← M
    no_improve ← 0
  else
    no_improve ← *no_improve* + 1
  end if
  if *no_improve* ≥ *patience* **then break**
end for
Return M^*

Algorithm 2 outlines a conservative fine-tuning strategy for limited-data, cross-hospital settings. Starting from a pre-trained OCR model, parameters are updated via mini-batch gradient descent on the source-hospital training set. Model performance is continuously assessed on a validation subset, with the optimal weights selected based on validation accuracy. Early stopping, governed by a predefined patience threshold, is employed to mitigate overfitting and ensure training stability. This strategy facilitates controlled adaptation while maintaining generalization in cross-hospital deployment settings and favors stable cross-domain transfer. $\eta = 10^{-5}$ favors stable cross-domain transfer, whereas $\eta = 10^{-4}$ speeds convergence but raises overfitting risk.

4) Computational complexity analysis

To better understand the computational implications of the proposed methods, the complexity of Algorithms 1 and 2 is briefly outlined below.

For evaluation (Algorithm 1), edit-distance–based metrics such as CER and NED require computing the distance between predicted and ground-truth sequences using dynamic programming. As a result, the evaluation cost increases with sequence length and dataset size but remains manageable for the datasets used in this study (≤3,294 samples, length ≤80).

For fine-tuning (Algorithm 2), the main cost comes from repeated forward and backward passes during training. The total cost depends on the number of epochs, dataset size, and model complexity. In practice, this varies significantly across OCR models due to differences in parameter size. PaddleOCR recognition models typically contain around 8 million parameters, while TrOCR variants range from approximately 62 to 345 million parameters, leading to notable differences in training and efficiency.

These results highlight a practical trade-off between model capacity and computational efficiency when deploying OCR systems in real-world hospital environments.

E. Evaluation Metrics

To comprehensively assess OCR performance under cross-hospital deployment conditions, both text detection and recognition metrics are employed. Detection metrics evaluate the accuracy of text region localization, while recognition metrics assess transcription correctness at the character and sequence levels. The evaluation procedure follows the protocol described in Algorithm 1.

1) Text detection metric

Text detection performance is evaluated using standard object localization metrics based on the overlap between predicted bounding boxes and ground-truth annotations. The Intersection over Union (IoU) between a predicted bounding box B_p and the ground-truth box B_g is defined as Ref. [16].

$$IoU = \frac{\left|B_p \cap B_g\right|}{\left|B_p \cup B_g\right|} \tag{2}$$

A detection is considered a true positive if $IoU \geq \theta$ where θ is the IoU threshold is (typically 0.50). Based on the numbers of True Positives (TP), False Positives (FP), and False Negatives (FN), the evaluation metrics are computed.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

The Harmonic mean (Hmean), equivalent to the F1-Score, is widely used in text detection benchmarks to summarize localization performance.

Detection performance is evaluated using precision–recall–based metrics. Specifically, it is summarized by the area under the precision–recall curve [27] at an IoU threshold of 0.50 and by mean Average Precision (mAP), where Average Precision (AP) represents the area under the precision–recall curve and mAP denotes the mean AP across classes. Localization robustness is further assessed across IoU thresholds from 0.50 to 0.95 in 0.05 increments. Metrics are reported separately for validation, test (in-domain), and unseen (cross-domain) sets to assess generalization.

2) Text recognition metric

Recognition performance is evaluated using sentence-level accuracy and edit-distance–based metrics. Sentence-level accuracy (ACC) measures the proportion of samples where the predicted text sequence exactly matches the ground-truth label. A prediction is considered correct only when the entire recognized string is identical to the reference; thus, ACC represents the ratio of correctly recognized samples to the total number of evaluated samples [28].

Character Error Rate (CER) measures character-level transcription errors using the Levenshtein edit distance between predicted and ground-truth sequences [8, 11].

$$CER = \frac{ED\left(y, \hat{y}\right)}{\left|y\right|} \tag{6}$$

where $ED(y, \hat{y})$ denotes the Levenshtein edit distance between the predicted sequence $\hat{y}$ and the ground truth $\hat{y}$ into y and $\left|y\right|$ is the ground truth sequence length. CER directly measures the proportion of character errors.

Normalized Edit Distance (NED) is also reported to provide a length-normalized similarity measure.

$$NED = 1 - \frac{ED\left(y, \hat{y}\right)}{\max\left(\left|y\right|, \left|\hat{y}\right|\right)} \tag{7}$$

where $ED(y, \hat{y})$ denotes the Levenshtein edit distance between the predicted sequence $\hat{y}$ and the ground truth y. The metric normalizes the edit distance by the maximum length of the two sequences, enabling fair comparison across different lengths, where higher values indicate better recognition performance [8, 11].

3) Thai word segmentation evaluation

Because Thai script lacks explicit word delimiters, accurate word boundary identification is essential for downstream NLP tasks [13, 29]. Let $W(s)$ denote a tokenizer that segments a sequence s into words; word-level errors are computed by comparing the tokenized OCR output with the tokenized ground truth:

$$WER = \frac{ED\left(W\left(y\right), W\left(\hat{y}\right)\right)}{\left|W\left(y\right)\right|} \tag{8}$$

where ED operates on word sequences rather than characters, and $\left|W(y)\right|$ denotes the number of words in the ground truth.

To assess word-level correctness, recognized text sequences are segmented using three Thai tokenization methods: newmm (dictionary-based maximum matching), deepcut (CNN–LSTM), and attacut (attention-based). Employing multiple tokenizers reduces tokenizer-specific bias and provides a more robust evaluation of word segmentation performance in pharmaceutical label recognition.

4) Paddle OCR-specific evaluation pipeline

PaddleOCR employs an advanced evaluation framework (postprocess_advanced) with Thai-specific post-processing, including character normalization and standardization, prior to metric computation. The enable_thai_wer option evaluates results across three tokenizers, providing a more comprehensive assessment beyond character-level accuracy.

5) Performance degradation metrics

To systematically quantify the impact of domain shift and enable fair comparison across architectures and training strategies, we introduce two complementary degradation metrics based on established cross-domain evaluation frameworks [29–31]. These metrics provide both absolute and relative assessments of performance degradation during deployment.

Absolute Performance Degradation (APD) measures the direct difference in accuracy between the in-domain test set and the cross-domain unseen set, calculated as Eq. (9):

$$\Delta_{abs} = P_{test} - P_{unseen} \tag{9}$$

where P_{test} is performance on the mixed test set (50% source, 50% other hospitals) and P_{unseen} is performance on the fully unseen set. The difference represents performance loss in percentage points; for instance, a 10-point decrease corresponds to a 10% accuracy drop when applied to unseen hospital domains.

Relative Performance Degradation (RPD) normalizes the absolute performance drop by the test performance, enabling fair comparison across models with different baseline capabilities, and is calculated as Eq. (10):

$$\Delta_{rel} = \frac{P_{test} - P_{unseen}}{P_{test}} \times 100\% \quad (10)$$

This normalization enables fair comparison across models with different baseline accuracies. For example, an 8-point drop corresponds to a 10% relative reduction for a model with 80% test accuracy but a 20% reduction for one with 40%, indicating greater vulnerability to domain shift despite the same absolute decrease. This reflects the principle that models with lower baseline performance tend to experience proportionally larger losses under distribution shift [32].

Absolute and relative drops capture different aspects of cross-domain robustness. Absolute drop reflects the actual accuracy loss during deployment, while relative drop measures proportional vulnerability to domain shift, enabling fair comparison across models with different baseline performance. A model with high test accuracy and a small relative drop indicates stronger generalization. Both metrics are reported separately for text detection and text recognition, enabling detailed analysis of domain shift effects on each component [6]. Following standard domain adaptation practice [33], validation-to-unseen comparisons measure overfitting to the source hospital, while test-to-unseen comparisons capture cross-domain degradation under partial domain overlaps. This dual-metric approach aligns with domain adaptation theory [21], where generalization depends on both source performance and source–target divergence, reflected by absolute and relative drops.

IV. Experiments and Results

This section compares text detection and recognition models for cross-hospital generalization. Results are organized to first assess text detection, followed by recognition performance, where the primary limitations of cross-domain generalization are observed.

A. Training Strategy for Text Detection Models

Text detection exhibits stable optimization and strong in-domain performance but limited predictive capacity for cross-domain generalization. Initialized from pre-trained weights [2, 3]. The PP-OCRv5 Mobile detector converges smoothly, with total loss decreasing from approximately 1.2–1.4 to 0.75–0.80 (Fig. 5). All loss components show consistent convergence, reflecting well-balanced multi-objective optimization.

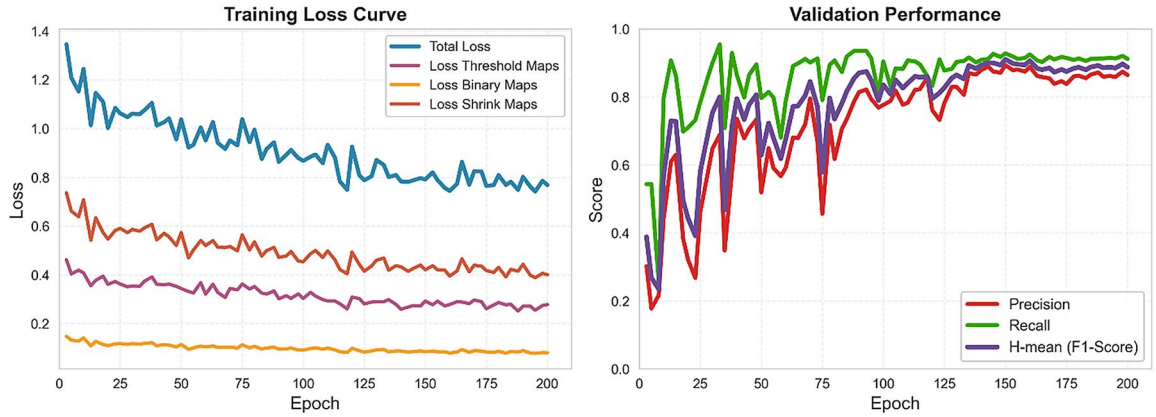


Fig. 5. PP-OCRv5 mobile detection training dynamics over 200 epochs.

Model Configurations:

- PP-OCRv5 Mobile (Pretrained): evaluated without fine-tuning to assess zero-shot transfer.
- PP-OCRv5 Mobile (Finetuned, *lr* = 0.001): fine-tuned on pharmaceutical labels, achieving optimal validation performance at epoch 150.

Validation performance improves rapidly in early epochs and stabilizes after around 50 epochs, achieving a Precision of 0.80–0.90, a Recall consistently above 0.90, and an H-mean between 0.88 and 0.92. The plateau observed near epoch 100 supports the use of early stopping at epoch 150 to prevent overfitting while preserving stable performance.

Detection maintains consistently high recall throughout training, indicating robust localization of text regions even in complex layouts. However, moderate precision suggests the presence of false positives, particularly in visually cluttered pharmaceutical labels. Importantly, strong validation results do not fully transfer to cross-hospital settings, indicating that detection is not the primary limitation under domain shift.

B. Text Recognition Models and Training

In contrast to detection, text recognition emerges as the primary factor limiting cross-domain OCR performance. A comparison between hybrid CNN–Transformer models (PaddleOCR) and fully Transformer-based models (TrOCR) reveals distinct generalization characteristics. Model specifications are summarized in Tables III and IV.

The lightweight PaddleOCR Mobile model [2, 3], despite its compact design (~8 M parameters), achieves competitive and more stable performance across domains. Its hybrid architecture, combining PPLCNetV3 feature extraction with SVTR-based sequence modeling [4], effectively balances local visual features and contextual dependencies. The use of CTC-based decoding further contributes to robustness under limited and noisy training data.

TABLE III. TEXT RECOGNITION MODEL ARCHITECTURE AND TRAINING CONFIGURATION

Model	Variant	Architecture	Decoder	Params	Optimizer	LR	Max Length
PaddleOCR	Mobile (v5)	SVTR	Multi-Head: CTC + NRTR	~8 M	Adam	1×10^{-5} / 1×10^{-4}	25 chars
TrOCR	thai-trocr	Vision Transformer	Electra Small (Thai)	~100 M	AdamW	1×10^{-5} / 1×10^{-4}	256 tokens
TrOCR	thai-trocr-thaigov-v2	Vision Transformer	WangchanBERTa Base	~200 M	AdamW	1×10^{-5} / 1×10^{-4}	256 tokens

TABLE IV. TRAINING CONVERGENCE AND BEST EPOCH SELECTION

Model	Learning Rate	Best Epoch	Val ACC	Val CER	Val NED
PaddleOCR Mobile	1×10^{-5}	80	0.9453	0.0089	0.9915
PaddleOCR Mobile	1×10^{-4}	100	0.9500	0.0076	0.9928
thai-trocr	1×10^{-5}	76	0.8953	0.0237	0.9522
thai-trocr	1×10^{-4}	67	0.9408	0.0117	0.9766
thai-trocr-thaigov-v2	1×10^{-5}	83	0.9469	0.0124	0.9861
thai-trocr-thaigov-v2	1×10^{-4}	85	0.9469	0.0124	0.9807

In contrast, Transformer-based TrOCR models [5, 10] exhibit higher capacity but reduced stability under cross-domain conditions. The openthaigpt/thai-trocr model (~100 M parameters) benefits from general Thai language pre-training, while kkatiz/thai-trocr-thaigov-v2 (~196 M parameters) demonstrates improved performance on domain-specific text due to synthetic government document pre-training [9]. However, both models rely on autoregressive decoding with beam search, making them more sensitive to distribution shifts and training data limitations.

The results indicate that larger Transformer-based models do not necessarily yield superior cross-domain performance. Instead, hybrid architecture offers a more favorable trade-off between accuracy, robustness, and computational efficiency, particularly in low-resource, real-world settings.

Overall, the results suggest that text detection is not the main bottleneck under domain shift, maintaining high recall (> 0.90) and stable H-mean (0.88–0.92). In contrast, recognition performance degrades more noticeably in cross-hospital scenarios, as evidenced by increased CER across models. Notably, lightweight hybrid CNN–Transformer models consistently yield lower CER and more stable performance than larger Transformer-based models, indicating that architectural design, rather than model scale alone, is critical for robust real-world OCR performance

Model Configurations and Training Protocol. Models are evaluated in base and fine-tuned settings (1×10^{-5}, 1×10^{-4}) under a unified PaddleOCR–TrOCR protocol to compare zero-shot transfer and limited-data adaptation. Fine-tuning runs up to 100 epochs with cosine scheduling and per-epoch validation. Models are selected by sentence-level ACC, with early stopping in Table IV and the best checkpoint used for test and unseen evaluation.

ACC represents sentence-level exact match accuracy, while lower CER and higher NED reflect improved character-level recognition. Although all models achieve high validation performance (89–95% ACC), indicating effective adaptation to the source hospital (Hospital A), this does not consistently generalize to cross-domain settings. For example, PaddleOCR ($lr = 1\times10^{-4}$) achieves the highest validation ACC (95.00%) with strong character-level metrics (CER = 0.0076, NED = 0.9928), whereas a more conservative setting ($lr = 1\times10^{-5}$) yields better unseen performance (80.47% vs. 73.83%) despite slightly lower validation ACC (94.53% vs. 95.00%). This discrepancy highlights a systematic validation-generalization gap, suggesting that in-domain metrics alone are insufficient for evaluating real-world OCR robustness.

Convergence patterns further reinforce this finding: PaddleOCR models converge later (epochs 80–100), indicating stable optimization under conservative learning rates, while TrOCR models converge earlier (epochs 67–85) but show greater susceptibility to overfitting source-domain patterns. Notably, Thai-trocr ($lr = 1\times10^{-5}$) attains lower validation ACC (89.53%) yet achieves the best cross-domain performance among TrOCR variants (51.67% unseen ACC). Overall, these results indicate that robust cross-domain OCR performance depends more on controlled optimization and architectural suitability than on validation accuracy or rapid convergence alone.

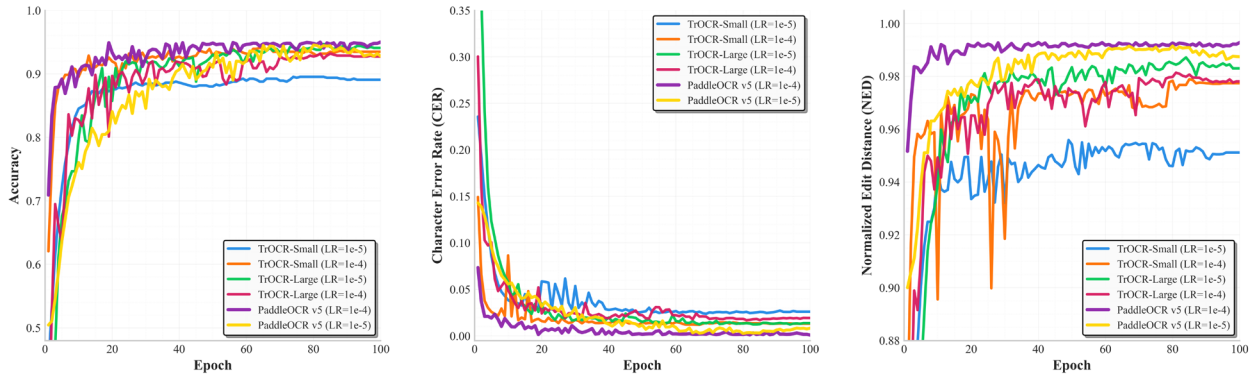
Training dynamics and convergence patterns: Fig. 6 shows validation performance over 100 epochs using Accuracy, CER, and NED, highlighting convergence behavior and learning rate sensitivity.

Accuracy (Fig. 6(a)): All models improve rapidly in the first 10–20 epochs. PaddleOCR converges smoothly after ~40 epochs (0.94–0.95), while TrOCR is more variable: Small converges faster but lower, and large reaches similar final accuracy. Higher learning rates speed up convergence but do not improve cross-domain performance.

CER (Fig. 6(b)): CER drops sharply by epoch 20. PaddleOCR achieves the lowest values (0.0076–0.0089), TrOCR-Small is higher and less stable, and TrOCR-Large is intermediate, reflecting stronger character-level supervision in PaddleOCR.

NED (Fig. 6(c)): NED rises quickly to 0.98–0.99. PaddleOCR performs best, followed by TrOCR-Large, while TrOCR-Small remains lower and less stable. Conservative TrOCR settings yield better cross-domain results despite weaker validation metrics.

Learning rate sensitivity: Validation metrics alone are insufficient, PaddleOCR remains stable across learning rates, but only conservative updates preserve generalization, while higher learning rates in TrOCR improve validation but increase overfitting, widening the validation–unseen gap.



(a) (b) (c)

Fig. 6. Recognition model training progression across 100 epochs. (a) Accuracy evolution. (b) CER evolution. (c) NED evolution.

C. Text Detection Performance on the Test Set

Text detection shows strong cross-domain stability (Table V), with pre-trained and fine-tuned PP-OCRv5 Mobile models achieving identical performance on both test (78.54% H-mean) and unseen sets (67.86% H-mean), despite large validation gains from fine-tuning (74.28% to 91.04%). This indicates that validation improvements do not translate to deployment and suggests that detection performance is largely governed by domain-invariant visual cues rather than domain-specific adaptation.

In contrast to the substantial degradation observed in text recognition, the relatively small drop from test to unseen (10.68 percentage points) highlights a clear asymmetry between localization and recognition. These results position text detection as a stable upstream component, while shifting the primary challenge of cross-hospital OCR to text recognition.

The fine-tuned model retains adequate performance on unseen data at moderate IoU thresholds (IoU@0.70: 63.01%; IoU@0.80: 53.06%), but drops sharply at IoU@0.90 (26.28%), reflecting challenges in precise boundary alignment across hospital layouts. In practice, IoU@0.70 is sufficient, as recognition tolerates moderately loose boxes (see in Table VI).

Detection shows a clear precision–recall trade-off: recall remains high across domains (91.80% test; 90.17% unseen), while precision declines (68.63% to 54.40%), indicating more false positives under domain shift. This recall-oriented behavior is preferable in OCR, as false positives can be filtered, whereas missed text cannot be recovered.

TABLE V. TEXT DETECTION PERFORMANCE AT IOU = 0.50

Model	Valid Hmean	Valid mAP	Test Hmean	Test Precision	Test Recall	Unseen Hmean	Unseen Precision	Unseen Recall
PP-OCRv5 Mobile (Pretrained)	0.7428	0.5658	0.7854	0.6863	0.9180	0.6786	0.5440	0.9017
PP-OCRv5 Mobile (FT lr = 0.001)	**0.9104**	**0.6701**	0.7854	0.6863	0.9180	0.6786	0.5440	0.9017

Note: Bold values indicate best performance; Hmean = Harmonic mean of Precision and Recall.

TABLE VI. DETECTION PERFORMANCE ACROSS IOU THRESHOLDS (FINE-TUNED MODEL)

Split	IoU@0.50	IoU@0.60	IoU@0.70	IoU@0.80	IoU@0.90	Time (ms)	FPS
Validation	**0.9104**	**0.8960**	**0.8396**	**0.6864**	0.2673	~759.88	~1.36
Test	0.7854	0.7798	0.7433	0.6367	0.3142	~759.88	~1.579
Unseen	0.6786	0.6556	0.6301	0.5306	0.2628	~669.79	~1.493

D. Text Recognition Performance on the Test Set

The experimental results presented in Table VII follow the three-level cross-hospital evaluation protocol described in Algorithm 1, assessing recognition performance under source-domain validation, mixed-domain testing, and fully unseen hospital conditions. This protocol enables a systematic evaluation of model generalization across heterogeneous hospital environments. To assess statistical reliability, 95% Confidence Intervals (CI) for CER and NED are estimated to use parametric bootstrap resampling (B = 10,000) on both the mixed-domain test set (n = 312) and the unseen-hospital set (n = 300). The resulting confidence intervals are reported in Table VII.

Table VII indicates that the PaddleOCR Mobile model fine-tuned with a conservative learning rate ($lr = 1\times10^{-5}$) achieves the best recognition performance on the unseen hospital dataset, reaching 80.47% accuracy and a CER of 0.0273. This result supports the conservative fine-tuning strategy described in Algorithm 2, where a smaller learning rate helps retain cross-domain transferability while adapting the model to source hospital data. In contrast, a more aggressive fine-tuning setting ($lr = 1\times10^{-4}$) converges more quickly but leads to a larger generalization gap between the test and unseen datasets.

In contrast, TrOCR models exhibit severe overfitting. Although validation accuracy remains high (>93%), performance drops sharply on the test set and deteriorates further on unseen data, suggesting limited cross-domain robustness even under partial domain overlap. Results

from base models further highlight differences in pre-training effectiveness. PaddleOCR demonstrates strong zero-shot transfer (65.62% on unseen data), outperforming all fine-tuned TrOCR variants, while TrOCR base models perform considerably worse, particularly the synthetic-heavy thai-trocr-thaigov-v2. This indicates that narrowly distributed or heavily synthetic pre-training data can negatively affect transferability.

In addition, the computational efficiency of the evaluated recognition architectures was examined to better understand their deployment implications. Table VIII summarizes the computational efficiency of the evaluated architecture. These results complement the performance analysis of Algorithm 1 by considering practical deployment factors, including inference latency, throughput, and model size. All measurements were obtained on a Google Colab Tesla T4 GPU (15 GB VRAM) using a batch size of one over a test set of 312 samples.

Table VIII shows the computational efficiency of the evaluated OCR architecture. The results indicate that PaddleOCR Mobile contains approximately 8 M parameters and achieves the lowest inference latency (47.83 ms per image) and the highest throughput (20.91 images per second). In contrast, the Transformer-based recognition models are substantially larger, with about 62 M parameters for thai-trocr and 345 M parameters for thai-trocr-thaigov-v2, resulting in higher inference latency and lower throughput.

TABLE VII. TEXT RECOGNITION PERFORMANCE WITH 95% CONFIDENCE INTERVALS

Model	Val ACC	Test ACC	Test CER [95% CI]	Test NED [95% CI]	Unseen ACC	Unseen CER [95% CI]	Unseen NED [95% CI]
PaddleOCR Base	-	0.6718	0.0600 [0.0585, 0.0615]	0.9404 [0.9389, 0.9419]	0.6562	0.0504 [0.0490, 0.0518]	0.9502 [0.9488, 0.9516]
PaddleOCR FT $lr = 1\times10^{-5}$	0.9453	**0.8086**	**0.0320** [0.0309, 0.0331]	**0.9686** [0.9675, 0.9697]	**0.8047**	**0.0273** [0.0263, 0.0284]	**0.9735** [0.9724, 0.9745]
PaddleOCR FT $lr = 1\times10^{-4}$	**0.9500**	0.8008	0.0357 [0.0345, 0.0369]	0.9647 [0.9635, 0.9659]	0.7383	0.0381 [0.0369, 0.0394]	0.9632 [0.9620, 0.9644]
thai-trocr Base	-	0.3173	0.2596 [0.2569, 0.2623]	0.7906 [0.7881, 0.7932]	0.2567	0.2852 [0.2822, 0.2881]	0.7773 [0.7746, 0.7800]
thai-trocr FT $lr = 1\times10^{-5}$	0.8953	0.5962	0.1023 [0.1004, 0.1042]	0.9120 [0.9102, 0.9138]	0.5167	0.1564 [0.1540, 0.1588]	0.8825 [0.8804, 0.8846]
thai-trocr FT $lr = 1\times10^{-4}$	0.9408	0.6026	0.1209 [0.1189, 0.1230]	0.9051 [0.9032, 0.9069]	0.4633	0.1895 [0.1869, 0.1921]	0.8608 [0.8586, 0.8630]
thaigov-v2 Base	-	0.0288	0.5885 [0.5854, 0.5916]	0.2713 [0.2685, 0.2740]	0.0500	0.5849 [0.5817, 0.5882]	0.3550 [0.3519, 0.3581]
thaigov-v2 FT $lr = 1\times10^{-5}$	0.9469	0.4872	0.2512 [0.2484, 0.2539]	0.7856 [0.7830, 0.7882]	0.3633	0.3641 [0.3610, 0.3672]	0.7087 [0.7057, 0.7116]
thaigov-v2 FT $lr = 1\times10^{-4}$	0.9317	0.4455	0.3516 [0.3486, 0.3546]	0.6974 [0.6945, 0.7004]	0.3067	0.4958 [0.4926, 0.4990]	0.5885 [0.5853, 0.5917]

Note: FT = Fine-tuned; Bold values indicate best performance within each metric category, *lr* = learning rate.

TABLE VIII. RECOGNITION MODEL EFFICIENCY COMPARISON

Model Architecture	Params (M)	Latency (ms/img)	Throughput (img/s)	Training Time (hr)
PaddleOCR Mobile	~8	47.83	20.91	-
thai-trocr	~62	57.08	17.52	2.11–2.18
thai-trocr-thaigov-v2	~345	60.68	16.48	2.31–2.38

E. Analysis of Domain Shift Impact

To quantify the domain shift effect comprehensively across both pipeline components. Table IX represents a systematic comparison of cross-domain robustness across detection and recognition models.

The results show an asymmetry in domain-shift sensitivity: detection degrades moderately (13.60%), while recognition varies more substantially (0.48%–31.15%) across models. Fine-tuning offers no cross-domain gain for detection, suggesting pre-trained models suffice, whereas recognition remains the main bottleneck. PaddleOCR with a conservative learning rate (1×10^{-5}) achieves strong robustness (<0.5% loss), while higher rates harm generalization despite similar validation accuracy.

TABLE IX. CROSS-DOMAIN PERFORMANCE DEGRADATION

Model	Component	Test Performance	Unseen Performance	Absolute Drop	Relative Drop
PP-OCRv5 Mobile (FT)	Detection	78.54%	67.86%	−10.68 pp	−13.60%
PaddleOCR Mobile ($lr = 1\times10^{-5}$)	Recognition	**80.86%**	**80.47%**	**−0.39 pp**	**−0.48%**
PaddleOCR Mobile ($lr = 1\times10^{-4}$)	Recognition	80.08%	73.83%	−6.25 pp	−7.81%
thai-trocr ($lr = 1\times10^{-5}$)	Recognition	59.62%	51.67%	−7.95 pp	−13.34%
thai-trocr-thaigov-v2 ($lr = 1\times10^{-4}$)	Recognition	44.55%	30.67%	−13.88 pp	−31.15%

Note: pp = percentage points; Relative drop = (Test−Unseen)/Test × 100%

TrOCR models show weaker cross-domain stability, even with conservative tuning, indicating that robustness depends more on transferable representations and controlled optimization than model scale. In practice, detection fine-tuning can be omitted, while recognition requires careful design and learning-rate control.

F. Zero-Shot Generalization to Unseen Pharmaceutical Label Domains

Zero-shot generalization is assessed by applying base OCR models trained on the source hospital directly to unseen pharmaceutical label images without further adaptation. Results on both in-domain test and unseen data reflect realistic deployment conditions in Table X.

TABLE X. BASE MODEL PERFORMANCE (ZERO-SHOT TRANSFER TO DRUG BAGS)

Model	Test Accuracy	Unseen Accuracy	Test-Unseen Gap
PaddleOCR Mobile (base)	**67.18%**	**65.62%**	−1.56 pp
thai-trocr (base)	31.73%	25.67%	−6.06 pp
thai-trocr-thaigov-v2 (base)	2.88%	5.00%	+2.12 pp

Performance varies markedly across models. PaddleOCR (base) achieves the most stable results (67.18% test; 65.62% unseen, −1.56 pp), indicating strong transferability. In contrast, thai-trocr (base) shows lower accuracy (31.73%; 25.67%, −6.06 pp), while thai-trocr-thaigov-v2 (base) performs poorly in both settings (2.88%; 5.00%), with the slight gain likely due to instability. Overall, zero-shot performance is strongly architecture-dependent, with PaddleOCR demonstrating superior cross-domain generalization.

To further examine this effect, Table XI summarizes the impact of learning rate on cross-domain performance. For PaddleOCR, a conservative learning rate (1×10^{-5}) achieves the highest unseen accuracy (80.47%) with a smaller validation–unseen gap (−14.06 pp), whereas a higher rate (1×10^{-4}) yields lower unseen accuracy (73.83%) and a larger gap (−21.17 pp), despite similar validation performance.

For Thai-trocr, fine-tuning improves validation accuracy (89.53–94.08%) and test performance (~60%), but generalization to unseen domains remains limited (46.33–51.67%), with large validation–unseen gaps (−37.86 to −47.75 pp). These results indicate that validation performance is not a reliable indicator of cross-domain readiness, particularly for Transformer-based models.

Implications for Model Design. Text recognition remains the main bottleneck under domain shift. Lightweight hybrid models (PaddleOCR) offer greater robustness and stability, whereas larger Transformer-based models are more sensitive to training dynamics. These results highlight that architectural suitability and controlled optimization are more important than model scale for reliable real-world OCR deployment.

TABLE XI. LEARNING RATE IMPACT ON KNOWLEDGE PRESERVATION AND GENERALIZATION

Model	LR	Val Acc	Test Acc	Unseen Acc	Val-to-Unseen Drop	Pre-train Gain
PaddleOCR	base	-	67.18%	65.62%	-	-
PaddleOCR	1×10^{-5}	94.53%	**80.86%**	**80.47%**	−14.06 pp	+14.85 pp
PaddleOCR	1×10^{-4}	**95.00%**	80.08%	73.83%	−21.17 pp	+8.21 pp
thai-trocr	base	-	31.73%	25.67%	-	-
thai-trocr	1×10^{-5}	89.53%	59.62%	51.67%	−37.86 pp	+26.00 pp
thai-trocr	1×10^{-4}	94.08%	60.26%	46.33%	−47.75 pp	+20.66 pp

V. DISCUSSION

The experimental results provide insights into OCR behavior under cross-hospital deployment conditions. By combining the structured evaluation protocol in Algorithm 1 with the fine-tuning strategy in Algorithm 2, this study enables systematic analysis of model generalization and training dynamics in limited-data medical environments.

The three-level evaluation framework allows examination of model performance under progressively increasing domain shift across hospitals. Although several models achieve high validation accuracy on the source-hospital dataset, their performance declines when evaluated on pharmaceutical labels from unseen hospitals, emphasizing the importance of evaluating OCR systems under realistic deployment conditions rather than relying solely on single-domain benchmarks.

Clear differences in cross-hospital generalization are observed among OCR architectures for Thai pharmaceutical labels. Fine-tuned PaddleOCR [3] achieves substantially higher accuracy on unseen hospital data (approximately 80%) than TrOCR variants [5] (approximately 30–50%), despite the latter having significantly more parameters. This gap also appears in the mixed-domain test set, indicating architectural differences in handling domain variation. Zero-shot evaluation further highlights the importance of pre-training quality, with PaddleOCR [3] achieving strong unseen performance without domain-specific fine-tuning (approximately 66%), surpassing fine-tuned TrOCR models [5].

Rather than attributing this behavior solely to a single cause, the results suggest that Transformer-based recognition models may exhibit greater sensitivity to domain shift under limited-data training conditions [5]. The three-level evaluation protocol therefore enables clear separation between in-domain validation performance and true cross-hospital generalization.

A. Component-Level Robustness and the Role of Pre-training

A clear asymmetry is observed between OCR components. Text detection [6] shows moderate and stable degradation (~14%), suggesting that transferable

localization features are largely captured during pre-training. In contrast, text recognition degrades substantially (up to > 30% in some models), identifying it as the primary bottleneck.

Recognition performance depends strongly on pre-trained representation quality. PaddleOCR (base) achieves 65.62% unseen accuracy without fine-tuning, while TrOCR base models reach only 5.00–25.67%, reflecting differences in pre-training diversity [2, 3, 5]. Synthetic pre-training further reveals a synthetic-to-real gap (2.88% unseen). Overall, recognition, not detection, is the main bottleneck in cross-hospital OCR.

B. *Error Characteristics under Domain Shift*

Qualitative observations indicate that performance degradation on unseen hospitals is associated with variations in typography, layout, and printing conditions, particularly affecting Thai tonal marks, diacritics, and character spacing. These variations, together with differences in formatting conventions and compound drug naming, introduce sequence-level inconsistencies in recognition. Consistent with the quantitative results, these findings confirm that text recognition remains the primary bottleneck, while detection is comparatively stable. A more systematic qualitative analysis of cross-hospital errors is therefore an important direction for future work. These issues intensify with incomplete local visual information, underscoring the need for robust feature extraction.

Critical detection errors often occur when bounding boxes fail to capture the full vertical extent of Thai characters, omitting tone marks or vowels (e.g., “ครั้ง”, “น้ำตาล”). Such truncation can propagate to recognition errors, particularly involving tone marks and diacritics, as illustrated in Fig. 7(a).

Character substitution among visually similar Thai letters (e.g., “ช” → “ข”) may also lead to clinically meaningful misinterpretations, such as “เช้า” (morning) being recognized as “เข้า” (enter) in Fig. 7(b).

Transformer-based models are more sensitive to crop distortions and missing local cues, while convolutional backbones provide more stable local feature extraction, based on observations from this dataset, as these conditions, though these findings are dataset-specific rather than general architectural limitations (see Fig. 8).

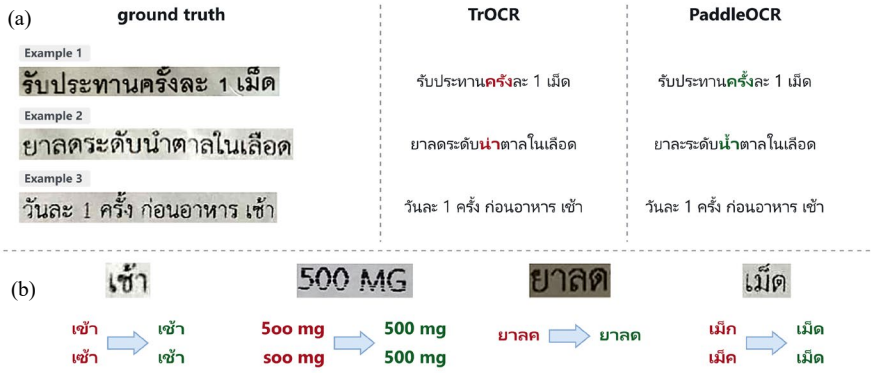


Fig. 7. OCR error examples in Thai medication instructions: (a) Thai tone mark and diacritic confusion across OCR models. (b) Clinically critical character substitution errors affecting dosage and timing.

Fig. 8. Impact of detection bounding box quality on recognition accuracy.

Additional errors include numerical confusion (e.g., “500” → “5oo”) and character merging from imperfect bounding boxes. These are highly sensitive to crop quality, where tight crops truncate content and loose crops introduce ambiguity. Recognition robustness thus depends on detection quality, representation learning, and decoding.

Transformer-based models are more sensitive to crop distortions and missing local cues, leading to higher error rates. In contrast, convolutional backbones provide more stable local feature extraction, though this observation is dataset-specific rather than a general limitation.

C. Architectural Inductive Biases and Fine-Tuning Dynamics under Limited Data

Recognition performance varies across architectures. PaddleOCR [3] achieves strong cross-hospital performance (~ 80% unseen), outperforming larger TrOCR models [5] (≈ 30–50%) despite having fewer parameters. Rather than attributing this behavior solely to overfitting, the results suggest that Transformer-based models may be more sensitive to domain shift under limited-data conditions, particularly when local visual cues are incomplete.

This highlights the role of architectural inductive bias. Hybrid CNN–Transformer models [3, 4], supported by CTC-based supervision [11], provide more stable character-level representations, whereas pure Transformer architectures [5, 12] depend more heavily on data diversity.

Learning rate analysis shows that conservative fine-tuning better preserves transferable representations and improves cross-domain robustness. In contrast, higher learning rates increase validation accuracy but decrease generalization, particularly for Transformer-based models. This indicates that validation performance alone is insufficient for assessing deployment readiness.

D. Decoder Effects, Decoding Strategy, and Practical Implications

Decoder pre-training differences may influence cross-domain performance. In the experiments, a smaller decoder trained on more diverse Thai corpora achieves higher unseen-hospital accuracy than a larger decoder trained mainly on formal administrative text, suggesting that linguistic diversity in pre-training may matter more than decoder capacity for pharmaceutical label recognition.

Decoding strategy appears to have a limited impact on cross-domain generalization. For example, TrOCR's autoregressive beam search does not substantially improve unseen-hospital performance compared with PaddleOCR's simpler decoding approach, indicating that decoding alone may have a limited effect under distribution shifts [5].

Overall, these results suggest that pre-training diversity should be prioritized over model scale when labeled data are limited, and synthetic data should be used cautiously unless it closely reflects real acquisition conditions. Architectures with stronger inductive biases, such as hybrid CNN–Transformer designs [3, 4], may provide more stable recognition in heterogeneous hospital environments. Moreover, validation accuracy alone may be insufficient for deployment decisions, as models with similar in-domain performance can behave differently on unseen hospital data.

E. Factors Influencing Cross-Hospital Generalization

Overall, cross-hospital generalization reflects the combined effects of component robustness, pre-training diversity, architectural inductive biases, and fine-tuning behavior. Text detection [6] remains relatively stable under domain shift, whereas recognition performance varies more with architecture and pre-training quality [3, 5]. Zero-shot results further suggest that domain-invariant representations are largely established during pre-training [2, 3, 5].

Architectural design also influences robustness under limited data. Hybrid CNN–Transformer architectures may provide stronger inductive biases that stabilize feature extraction across heterogeneous layouts [3, 4], while the larger validation-to-unseen gaps observed for TrOCR models [5, 12] indicate greater sensitivity to domain variation when training data diversity is limited. These findings suggest that increasing model scale alone does not guarantee better cross-hospital generalization.

Common failure modes include vowel and diacritic omissions in low-contrast labels, inconsistent handling of numeric–text boundaries across templates, and recognition errors with unfamiliar fonts or irregular character spacing.

F. Security and Reliability Considerations

From a deployment perspective, OCR errors may pose safety risks in medical contexts. Misrecognition of dosage, timing, or drug names can lead to clinically significant misunderstandings. While this study focuses on model performance, the findings indicate that robustness to domain shift is critical for system reliability. In practice, systems should incorporate confidence-based filtering and human-in-the-loop verification for high-risk cases.

G. Practical Implications for OCR Deployment

The findings offer practical guidance for cross-hospital OCR deployment. Detection can be used without fine-tuning, reducing annotation cost and simplifying deployment, while recognition requires careful architectural choice and controlled optimization to maintain robustness.

H. Limitations and Future Scopes

First, the study is constrained by limited single-hospital data (3,294 recognition; 403 detection), where pre-training dominates performance and validation accuracy is insufficient for model selection; conservative optimization and architectural bias act as implicit regularizers.

Second, the dataset is small and geographically limited (±2–3 recognition, ±5–7 detection uncertainty), and the analysis focuses on quantitative results without detailed error annotation.

Third, practical deployment factors such as privacy, security, and system integration are not addressed. Future work will extend to multi-institution datasets, systematic error analysis, and privacy-aware OCR with domain-aware synthesis, active learning, and multi-source training.

VI. Conclusion

This study presents a systematic evaluation of OCR systems for Thai pharmaceutical label recognition under realistic cross-hospital deployment conditions. A structured three-level evaluation framework is proposed to analyze model behavior under progressively increasing domain shift, offering a more reliable assessment of generalization than conventional single-domain evaluation.

The results reveal a clear asymmetry between OCR components. Text detection remains relatively stable across hospitals, suggesting that transferable localization knowledge is largely captured during pre-training. In contrast, text recognition emerges as the primary bottleneck, exhibiting substantially greater performance degradation under domain shifts.

Recognition robustness is strongly influenced by pre-training diversity and architectural design. Hybrid CNN–Transformer models achieve superior cross-domain performance with significantly fewer parameters than large Transformer-based models. Conservative fine-tuning further enhances robustness by preserving transferable representations while reducing sensitivity to source-domain bias.

From a practical standpoint, detection fine-tuning can be omitted, simplifying deployment, whereas recognition requires careful model selection and controlled training. The proposed evaluation framework provides a systematic approach for distinguishing in-domain performance from true cross-domain generalization.

Qualitative observations show that cross-hospital differences in typography, layout, and printing conditions lead to systematic recognition errors, especially in Thai tonal marks, diacritics, and character spacing. These errors are exacerbated by imperfect detection crops and the structural sensitivity of Thai script, resulting in sequence-level inconsistencies under domain shift. Consistent with quantitative results, recognition remains the main bottleneck, while detection is relatively stable, highlighting the need for more robust character-level representations and further analysis of cross-domain errors.

Overall, this work demonstrates that OCR robustness in cross-hospital environments depends more on architectural design, pre-training diversity, and optimization strategy than on model scale alone. Future work will extend this study to larger multi-institutional datasets, more detailed error analysis, and privacy-aware OCR deployment in clinical systems.

APPENDIX: LIST OF NOTATION

Table AI is the List of Notations used in the proposed OCR evaluation framework. The principal symbols and variables used in Algorithms and evaluation procedures throughout this study.

TABLE AI. NOTATIONS

Symbol	Description
D_A	Labeled dataset collected from the source hospital
D_{train}	Training subset sampled from the source-hospital dataset
D_{val}	Validation subset sampled from the source-hospital dataset
D_{mix}	Mixed-domain dataset containing samples from both source and other hospitals
D_{unseen}	Dataset collected from previously unseen hospitals used for cross-hospital evaluation
M	OCR model used for training and inference
M_0	Pre-trained OCR model initialization
M^*	Fine-tuned OCR model with the best validation performance
E	Number of training epochs
E_{max}	Maximum number of training epochs
$epoch$	Current training epoch
N	Number of samples in the dataset
B	Mini-batch sampled from the training dataset
η	Learning rate used during model training
y	Ground-truth text sequence
$\hat{y}$	Predicted text sequence generated by the OCR model
L	Loss function used during training
P_{val}	Validation performance score
P_{test}	Mixed-domain test performance score
P_{unseen}	Performance score on unseen hospital data
val_acc	Validation accuracy measured during training
$best\_val\_acc$	Highest validation accuracy observed during training
$no\_improve$	Counter for epochs without validation improvement
$patience$	Early stopping threshold for training termination
val_gap	Difference between validation and unseen performance
Δ_{abs}	Absolute performance difference between test and unseen datasets
Δ_{rel}	Relative performance difference between test and unseen datasets

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

W.S.N. conceptualized the research idea, designed the methodology, developed the experimental framework, conducted the experiments and formal analysis, curated the data, prepared the visualizations, and drafted the initial manuscript. P.J. contributed to methodology, validation, formal analysis, and manuscript revision. Ku.S.N. and Ka.S.N. assisted with data curation, investigation, validation, and visualization. P.N. supervised the study, contributed conceptualization and methodology, provided resources, administered the project, acquired funding, and critically revised the manuscript. All authors discussed the results and approved the final manuscript for publication.

FUNDING

This research was funded by the Thailand Science Research and Innovation Fund (Fundamental Fund 2026), and the University of Phayao.

ACKNOWLEDGMENT

This research project also received support from many advisors, academics, researchers, students, and staff. The authors thank everyone for their support and cooperation in completing this research.

REFERENCES

[1] W. S. Nuankaew, F. Jailangka, A. Khampraphai, A. Kamka, P. Nasa-Ngiu, and P. Nuankaew, "AI for medical informatics: The application of optical character recognition technology in the transcription of pharmaceutical labels," *Journal of Advances in Information Technology*, vol. 16, no. 9, pp. 1338–1351, 2025. doi: 10.12720/jait.16.9.1338-1351

[2] Y. Du *et al.*, "PP-OCR: A practical ultra lightweight OCR system," arXiv preprint, arXiv:2009.09941, 2020.

[3] C. Li *et al.*, "PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system," arXiv preprint, arXiv:2206.03001, 2022.

[4] Y. Du *et al.*, "SVTR: Scene text recognition with a single visual model," arXiv preprint, arXiv:2205.00159, 2022.
[5] M. Li *et al.*, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 11, pp. 13094–13102. doi: 10.1609/aaai.v37i11.26538
[6] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," arXiv preprint, arXiv:1911.08947, 2019.
[7] P. Nitayavardhana *et al.*, "Streamlining data recording through optical character recognition: A prospective multi-center study in intensive care units," *Crit Care*, vol. 29, no. 1, 117, 2025. doi: 10.1186/s13054-025-05347-1
[8] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.
[9] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2023.
[10] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, 2015.
[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376. doi: 10.1145/1143844.1143891
[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, 2021.
[13] A. Marzal and E. Vidal, "Computation of normalized edit distance and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 926–932, Sep. 1993. doi: 10.1109/34.232078
[14] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," arXiv preprint, arXiv:2106.08254, 2022.
[15] C. Cui *et al.*, "PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model," arXiv preprint, arXiv:2510.14528, 2025.
[16] E. Hsu, I. Malagaris, Y. F. Kuo, R. Sultana, and K. Roberts, "Deep learning-based NLP data pipeline for EHR-scanned document information extraction," *Jamia Open*, vol. 5, no. 2, ooac045, 2022. doi: 10.1093/jamiaopen/ooac045
[17] L. G. Singh and S. E. Middleton, "Tabular context-aware optical character recognition and tabular data reconstruction for historical records," *International Journal on Document Analysis and Recognition*, vol. 28, no. 3, pp. 357–376, 2025. doi: 10.1007/s10032-025-00543-9
[18] X. Liu, J. Meehan, W. Tong, L. Wu, X. Xu, and J. Xu, "DLI-IT: A deep learning approach to drug label identification through image and text embedding," *BMC Med Inform Decis Mak*, vol. 20, no. 1, 68, 2020. doi: 10.1186/s12911-020-1078-3
[19] C. Cui *et al.*, "PaddleOCR 3.0 Technical Report," arXiv preprint, arXiv:2507.05595, 2025.
[20] D. Dhinakaran, L. Srinivasan, S. U. Sankar, and D. Selvaraj, "Quantum-based privacy-preserving techniques for secure and trustworthy internet of medical things an extensive analysis," *Quantum Inf. Comput.*, vol. 24, no. 3 & 4, pp. 227–266, 2024.
[21] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach Learn*, vol. 79, no. 1, pp. 151–175, May 2010, doi: 10.1007/s10994-009-5152-4
[22] P. W. Koh *et al.*, "WILDS: A benchmark of in-the-wild distribution shifts," arXiv preprint, arXiv:2012.07421, 2021.
[23] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012. doi: 10.1016/j.patcog.2011.06.019
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90
[25] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, Jan. 1974. doi: 10.1145/321796.321811
[26] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, Jan. 2018. doi: 10.1137/16M1080173
[27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," *Int J Comput Vis*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4
[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009. doi: 10.1016/j.ipm.2009.03.002
[29] E. Vidal, A. Marzal, and P. Aibar, "Fast computation of normalized edit distances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 9, pp. 899–902, Sep. 1995. doi: 10.1109/34.406656
[30] M. W. Ma *et al.*, "Extracting laboratory test information from paper-based reports," *BMC Med Inform Decis Mak*, vol. 23, no. 1, 251, Nov. 2023. doi: 10.1186/s12911-023-02346-6
[31] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," arXiv preprint, arXiv:2012.12877, 2021.
[32] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, 2020. doi: 10.1145/3400066
[33] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1180–1189.