Specialization Over Scale: Compact Open-Weight VLMs for Medical OCR via Low-Rank Adaptation

Anonymous ACL submission

Abstract

Optical character recognition (OCR) underpins many clinical workflows, yet cost and data-governance constraints discourage reliance on large proprietary vision-language models (VLMs). We investigate whether compact open-weight VLMs can be adapted via Low-Rank Adaptation (LoRA) into competitive medical OCR systems that deliver clinical value. Using a controlled study within the Gemma model family, we isolate the effects of model scale, domain alignment, and fine-tuning data composition on OCR performance. Specifically, we construct a corpus spanning general and medical domains across text extraction (TE) and key information extraction (KIE) tasks, and apply staged LoRA adaptation to Gemma 4B and MedGemma 4B, using Gemma 27B as a zero-shot reference. Our fine-tuned MedGemma 4B matches or surpasses the 27B model on most targeted OCR benchmarks, albeit at the cost of out-of-distribution generality. Compared to Gemma 4B, MedGemma shows stronger in-domain zero-shot performance and larger fine-tuning gains. As an OCR extractor in a downstream Protected Health Information (PHI) identification pipeline, MedGemma 4B outperforms EasyOCR and Gemini 2.5 Flash and reduces missed PHI instances by 80% relative to its pre-trained version. Our results show that compact open-weight VLMs can be a promising alternative to large models in targeted, privacy-sensitive clinical OCR settings.

1 Introduction

Optical character recognition (OCR) is a critical component of many clinical workflows, including transcription of handwritten and printed clinical notes (Kumar et al., 2023), extraction of structured information from medical forms (Ponnuru et al., 2024), insurance claims and billing validation (Lerisha J et al., 2025), and automated identification of Protected Health Information (PHI) in medical images (Truong et al., 2025b). Medical OCR is further challenged by handwriting, low-resolution scans, inconsistent fonts, abbreviations, and subtle imprints (Liu et al., 2024)—conditions under which traditional OCR systems are known to underperform (Patel, 2025; Jain et al., 2021; Tabbone and Pelingre, 2025).

In contrast to traditional engines, vision-language models (VLMs) can jointly process visual and textual information, enabling semantic reasoning, structured output generation, and instruction-following behavior (Li et al., 2025; Gupta et al., 2024). Recent studies have reported strong performance for VLMs on OCR-related tasks (Nagaonkar et al., 2025; Greif et al., 2025; Chen et al., 2025). However, many of the strongest VLMs to date are large and proprietary, which can limit their suitability in privacy-sensitive domains. In healthcare settings, reliance on external APIs raises data-governance and compliance concerns and introduces deployment dependencies through vendor-controlled model updates.

Compact open-weight VLMs offer a promising alternative. They can be adapted under realistic compute budgets, deployed with greater operational control, and reduce reliance on external OCR services. These properties make them attractive components for privacy-sensitive pipelines. However, in their pre-trained state, compact models typically underperform larger versions (Kaplan et al., 2020). At the same time, targeted design decisions and thoughtful architectural engineering can compensate for scale differences (Sinha et al., 2025; He et al., 2024; Jin et al., 2025). The extent to which compact open VLMs can approach larger-model OCR performance through parameter-efficient fine-tuning (PEFT) without architectural modifications, particularly in medical settings, remains unexplored.

It is also unclear whether domain-aligned pre-training provides an advantage over general-purpose initialization for OCR adaptation, as ex-

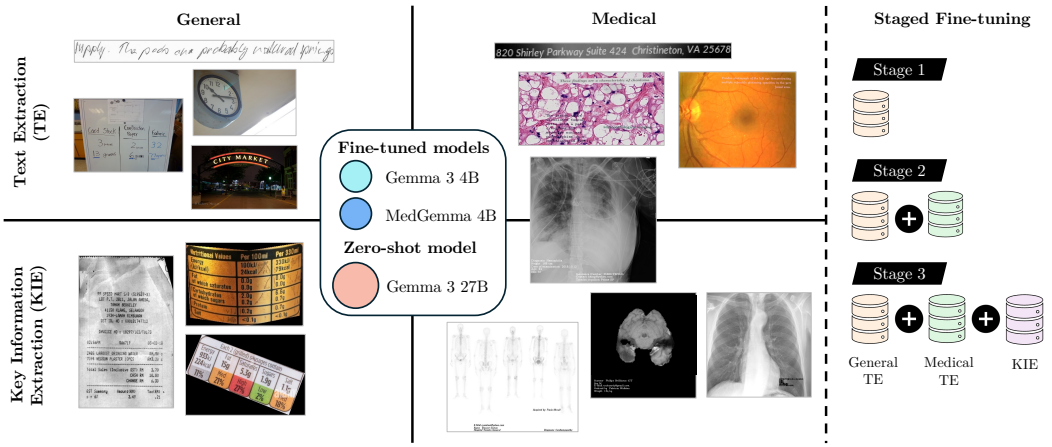

Figure 1: Overview of the experimental framework and staged fine-tuning strategy. The curated data corpus spans two domains (general and medical) and two OCR tasks (text extraction and key information extraction). Compact models (Gemma 3 4B and MedGemma 4B) are fine-tuned with PEFT (LoRA) on progressively more extensive data and task combinations to isolate the effects of each increment. The 6.75 times larger Gemma 3 27B model serves as a zero-shot baseline for comparison.

isting evidence is inconclusive (Jeong et al., 2025; Zhou et al., 2025). Finally, OCR is typically evaluated in isolation using transcription metrics, without assessing whether OCR-level gains translate into downstream clinical utility.

We study these problems through three main research questions: how compact medical and general-purpose open VLMs compare with a much larger VLM across text extraction (TE) and key information extraction (KIE) OCR tasks in general and medical domains; how PEFT, specifically Low-Rank Adaptation (LoRA) (Hu et al., 2021), changes this picture; and whether OCR gains translate into improvements in a downstream task. Specifically, we benchmark MedGemma 4B (Sellergren et al., 2025), Gemma 3 4B, and Gemma 3 27B (Gemma Team et al., 2025) across a curated corpus spanning two OCR task abstraction levels—lower-level TE (raw text transcription) and higher-level KIE (structured semantic parsing)—and two domains—general and medical. As shown in Figure 1, we fine-tune the two 4B models with LoRA, using Gemma 27B as a large zero-shot reference. The staged adaptation design employs freshly initialized adapters trained on a progressively expanded corpus, which allows for a controlled study, isolating the effects of model scale, domain alignment, and training data composition. The best-performing fine-tuned model is evaluated on a downstream PHI identification task.

These experiments yield several practically relevant findings. Fine-tuned MedGemma 4B bridges the gap with the pre-trained Gemma 3 27B on targeted OCR benchmarks at the expense of out-of-distribution generalization. Domain-aligned pre-training enables competitive zero-shot in-domain performance and amplified fine-tuning gains. Single-task fine-tuning induces *output format overfitting*, where outputs collapse to the training format despite explicit prompt instructions. Fine-tuning also reduces hallucination and improves downstream reliability. Overall, the results point to the viability of deploying small models for privacy-preserving reasons in targeted domains. Our key contributions are:

- We show that a compact open-weight medical VLM fine-tuned with LoRA can match or outperform a much larger untuned generalist on targeted OCR tasks with a clear specialization-generalization trade-off and findings on domain alignment, format overfitting, and hallucination reduction described above.
- We demonstrate the downstream value of these OCR gains in a PHI identification pipeline, where replacing the OCR component with the fine-tuned compact model reduces missed PHI instances relative to traditional OCR and large proprietary VLM baselines.
- We release fine-tuned model weights, OCR data preprocessing and synthetic generation pipelines, and LoRA fine-tuning and benchmarking framework to support future research in locally deployable OCR systems.[1]

[1] Upon acceptance, the resources will be publicly accessi-

2 Related Work

Traditional OCR engines perform well on clean printed text but often struggle with noisy layouts, handwriting, and domain-specific annotations (Jain et al., 2021; Tabbone and Pelingre, 2025). Recent general-purpose VLMs consistently outperform them on heterogeneous inputs (Nagaonkar et al., 2025; Greif et al., 2025; Singh et al., 2025), with benchmarks such as CC-OCR (Yang et al., 2024) and OCRBench v2 (Fu et al., 2025) confirming broader OCR competence. Open-weight models are also increasingly competitive with proprietary systems on specialized tasks (Fu et al., 2025; Wang et al., 2025), making them attractive for healthcare deployment because they support local hosting, version control, and auditability without external data transmission (Alassan et al., 2024; Hanke et al., 2024). However, it remains unclear to what extent compact open-weight VLMs, once fine-tuned, can compete with much larger zero-shot models on OCR tasks in medical settings.

PEFT methods, including LoRA (Hu et al., 2021), are widely used to adapt compact models under constrained compute budgets and can make specialization feasible without full model retraining (He et al., 2024; Chen et al., 2024). Despite this, LoRA-based fine-tuning for medical OCR remains underexplored, and the effects of fine-tuning data composition on OCR behavior have not been systematically characterized. Evidence on whether medical pre-training benefits downstream fine-tuning is mixed. Jeong et al. (2025) report limited consistent gains on medical QA, Yang et al. (2024) further note that medical VLMs can exhibit weaker instruction-following, while Zhou et al. (2025) document improved clinical generalization and LLaVA-Med substantially outperforms its base model on biomedical benchmarks (Li et al., 2023). In the context of medical and general OCR, it remains unclear whether domain-aligned VLMs offer inherent advantages or whether solely task-specific adaptation is relevant for the OCR performance.

Most OCR fine-tuning studies evaluate transcription in isolation, using standalone extraction metrics rather than downstream task performance. In practice, however, OCR often serves as one component in larger clinical pipelines, such as PHI identification (Truong et al., 2025a,b). Whether improvements in OCR quality translate into measurable gains in such settings remains largely untested.

Our work addresses this gap through a controlled comparison within one model family, staged LoRA adaptation, and evaluation across domain, task level, and downstream utility.

3 Methods

Our experimental design isolates three factors: model scale, domain-aligned pre-training, and fine-tuning data composition, and evaluates their impact on OCR performance and downstream utility.

3.1 Models

We select models within the Gemma family (Gemma Team et al., 2025) to control for architecture while varying scale and domain alignment under realistic inference and fine-tuning constraints. Specifically, we use Gemma 3 4B as a compact general-purpose VLM, MedGemma 4B (Sellergren et al., 2025) as its domain-specialized medical counterpart at matched scale, and Gemma 3 27B as a larger general-purpose zero-shot reference. The 27B model is run with 4-bit quantization. Hereafter, we omit the version number (3) for brevity. We additionally include EasyOCR (Jaided AI, 2020) as a traditional OCR baseline, motivated by its use in the PHI pipeline (Truong et al., 2025b).

3.2 OCR Corpus

We curate a standardized corpus along two axes: domain (general, medical) and task (lower-level text extraction, higher-level key information extraction), enabling controlled analysis of domain and task composition effects. As annotated medical OCR data is scarce, the medical portion combines existing datasets with synthetic generation. The full dataset descriptions and construction details are provided in Appendix A.1–A.5.

All datasets are converted to a unified JSON annotation schema, preprocessed to improve label consistency, and split into train/validation/test partitions with leakage prevention where relevant. To isolate domain effects from data volume, the general and medical TE training sets are balanced to equal size. Samples with ground truths longer than 512 tokens are removed to match the inference limit. The final corpus contains around 125k general-domain TE samples (comprising IAM-line (Marti and Bunke, 2002), TotalText (Chng and Chan, 2017), TextOCR (Singh et al., 2021), and SynthText (Gupta et al., 2016)), 129k medical-domain TE samples (MultiCaRe-TE (Nievas Offi-

ble; the code will be distributed under Apache 2.0 License.

dani, 2024), PHI-TE (Truong et al., 2025a), and TE-synthetic), and 11k KIE samples spanning SROIE (Huang et al., 2019), POIE (Kuang et al., 2023), and PHI-KIE (Truong et al., 2025a) (Table 5).

3.3 Fine-Tuning Strategy

Training Data	S1	S2	S3
General TE TextOCR, TotalText, IAM-line, SynthText	✓	✓	✓
Medical TE MultiCaRe-TE, PHI-TE, TE-synthetic		✓	✓
TE + KIE SROIE, POIE, PHI-KIE			✓

Table 1: Progressive LoRA fine-tuning setup for Gemma 4B and MedGemma 4B. Each stage (S1–S3) trains newly initialized adapters on top of the original pre-trained model checkpoints.

We use LoRA (Hu et al., 2021) for parameter-efficient specialization of the compact models. We use a three-stage LoRA setup in which each stage trains freshly initialized adapters on a progressively expanded corpus (Table 1). This design isolates the effects of medical-domain exposure and multi-task training.

Stage 1 fine-tunes both compact VLMs on general-domain TE only.

Stage 2 fine-tunes both models on combined general and medical TE data.

Stage 3 fine-tunes both models on the full corpus, adding KIE to assess multi-task effects.

Hyperparameters are selected by validation on Stage-1 data; we use rank $r = 64$, scaling factor $\alpha = 128$, AdamW, cosine decay, a learning rate of 3×10^{-5}, and 5% warmup. Full tuning details are provided in Appendix B.2 and B.3. The fine-tuning experiments are conducted on NVIDIA A100 (40GB) GPUs, requiring ∼207 wall-clock hours per model (spanning 38.8h, 93.5h, and 74.7h for Stages 1–3, respectively).

3.4 Inference

For TE, a single rule-based prompt is used across all datasets, instructing the model to transcribe all visible text exactly as it appears. For KIE, dataset-specific prompts include a full JSON schema listing all possible keys; the model is instructed to fill present values and return empty strings for absent keys. This open-schema design better reflects real deployment scenarios where field presence cannot be assumed, unlike benchmarks that provide only the keys known to be present. The exact prompts are provided in Appendix B.1. Maximum generation length is set to 512 tokens. A post-processing pipeline repairs corrupted JSON responses (e.g., markdown fences, unbalanced braces, null value variants) prior to evaluation.

3.5 Evaluation

Curated corpus. All models are evaluated on held-out test splits before fine-tuning and after each stage. For TE, the primary metric is mean token-level F1 score, which captures transcription overlap while remaining robust to moderate formatting differences; Character Error Rate (CER) and Word Error Rate (WER) serve as complementary error-sensitive metrics. For KIE, macro token-level F1 is the primary metric, complemented by Average Normalized Levenshtein Similarity (ANLS) and binary F1 for structural correctness. CER and WER are computed using trimmed means (10% outlier rate) to prevent hallucinated repetitive generations from dominating aggregate scores; severe hallucination behavior is analyzed separately through untrimmed error statistics. Ground truths and predictions undergo dataset-specific normalization before metric computation.

Generalization. We evaluate the best fine-tuned model, its pre-trained version, and Gemma 27B on the CC-OCR benchmark (Yang et al., 2024). Datasets are partitioned into those qualitatively similar to the fine-tuning corpus (synthetic user-generated content, scene and document text) and those that diverge substantially (e.g., multilingual text, mathematical content, tabular data), enabling separate assessment of fine-tuning gains and generalization costs.

Downstream PHI pipeline. Following Truong et al. (2025b), we evaluate the best fine-tuned model within a downstream PHI identification pipeline using the RadPHI-test dataset (Truong and Lenga, 2025), which contains simulated PHI imprints from 16 categories overlaid on medical images. We replace the OCR component of the original pipeline with the evaluated model, while retaining YOLO-based text localization and Gemini 2.5 Flash for PHI classification. Pipeline evaluation prioritizes recall and false negatives as the primary safety metrics, since missed PHI constitutes a privacy violation in clinical de-identification.

4 Results and Analysis

4.1 Pre-Trained Baselines: Scale vs. Domain Alignment

As shown in Table 2, all three VLMs substantially outperform EasyOCR across TE datasets, often by large margins, confirming the limitations of traditional OCR when processing heterogeneous inputs such as handwriting and low-resolution images. Among VLMs, model scale is the strongest predictor of zero-shot performance: Gemma 27B achieves the highest mean F1 on six of seven TE datasets and all three KIE datasets, reaching 90.41% on PHI-TE, 83.35% on MultiCaRe-TE, and 90.18% on SROIE. Gemma 4B follows consistently below the 27B model, with the gap most pronounced on datasets with complex or heterogeneous text, such as TextOCR and MultiCaRe-TE.

MedGemma 4B partially compensates for scale within the medical TE domain. Despite being 6.75 times smaller, it matches or exceeds Gemma 27B on PHI-TE dataset, while underperforming on general-domain datasets. This pattern is consistent with domain-aligned pre-training improving in-domain OCR readiness at the cost of broader zero-shot coverage. However, the same advantage does not extend to higher-level OCR performance in the KIE: despite strong PHI-TE performance, MedGemma ranks lowest on PHI-KIE, indicating that medical TE skill does not directly transfer to in-domain structured extraction without task-specific adaptation. MedGemma's ANLS remains more competitive than its macro F1 on KIE task (Appendix C.1), suggesting that its KIE errors reflect failures of field completion more than uniformly weak text recognition.

Inference latency further favors compact models. Gemma 27B requires $\sim$28 seconds per sample on KIE compared to $\sim$5 seconds for 4B models, and $\sim$9 vs. $\sim$4 seconds on TE, highlighting the practical cost of scale.

4.2 PEFT Closes and Reverses the Scale Gap

LoRA fine-tuning substantially improves both small VLMs across tasks (Table 2). By Stage 3, MedGemma 4B surpasses Gemma 27B on six of seven TE datasets and two of three KIE datasets, reaching 94.51% macro F1 on SROIE and 96.18% on PHI-KIE, compared to 90.18% and 86.18% for the large baseline (Figure 2). Stage-3 Gemma 4B also exceeds Gemma 27B on six of seven TE datasets, but MedGemma 4B remains the strongest of the two fine-tuned models across all benchmarks.

MedGemma 4B benefits more consistently from PEFT across stages, including general-domain datasets where medical pre-training provides no obvious a priori advantage. Stage-1 gains on TextOCR are 51% for MedGemma 4B vs. 19% for Gemma 4B; on SynthText, 54% vs. 20%. This pattern is consistent with domain-aligned pre-training providing a more effective initialization for OCR adaptation in this model family—possibly through representational priors that support fine-grained visual text processing, although causal attribution remains limited by the undisclosed composition of the pre-training corpus.

The staged design further shows that data composition matters. Stage 1 already yields large TE gains, while Stage 2 adds substantial improvement on medical TE with only modest losses on some general datasets, reducing the gap in the domain between small general and medical models. Stage 3 restores structured extraction performance with a slight improvement across the TE datasets, suggesting that text recognition and structured reasoning benefited from mutual reinforcement, although the relatively small size of the KIE data impedes interpretation. Per-dataset dynamics across stages are detailed in Appendix C.2. Taken together, these results show that in targeted settings compact open-weight VLMs can close or reverse the scale gap to a much larger zero-shot model.

4.3 Output Format Overfitting

Stage-1 fine-tuning on plain-text TE alone causes failure of structured-output capability: KIE macro F1 drops near zero for both models despite improved TE. Qualitative inspection suggests that the underlying recognition capability is preserved, but models return text as plain strings, disregarding the JSON schema requested in the KIE prompt. We term this *output format overfitting*: the model learns to prefer the format encountered during training and applies it unconditionally, overriding explicit prompt instructions at inference time.

Stage 2 yields a slight recovery without any KIE supervision. The structural variation in medical TE data, particularly label-value imprint pairs resembling key-value fields, might inadvertently reinforce output structuring. Full remediation is achieved at Stage 3 through explicit KIE supervision: even adding comparatively limited KIE data allows the models to recover and surpass the pre-

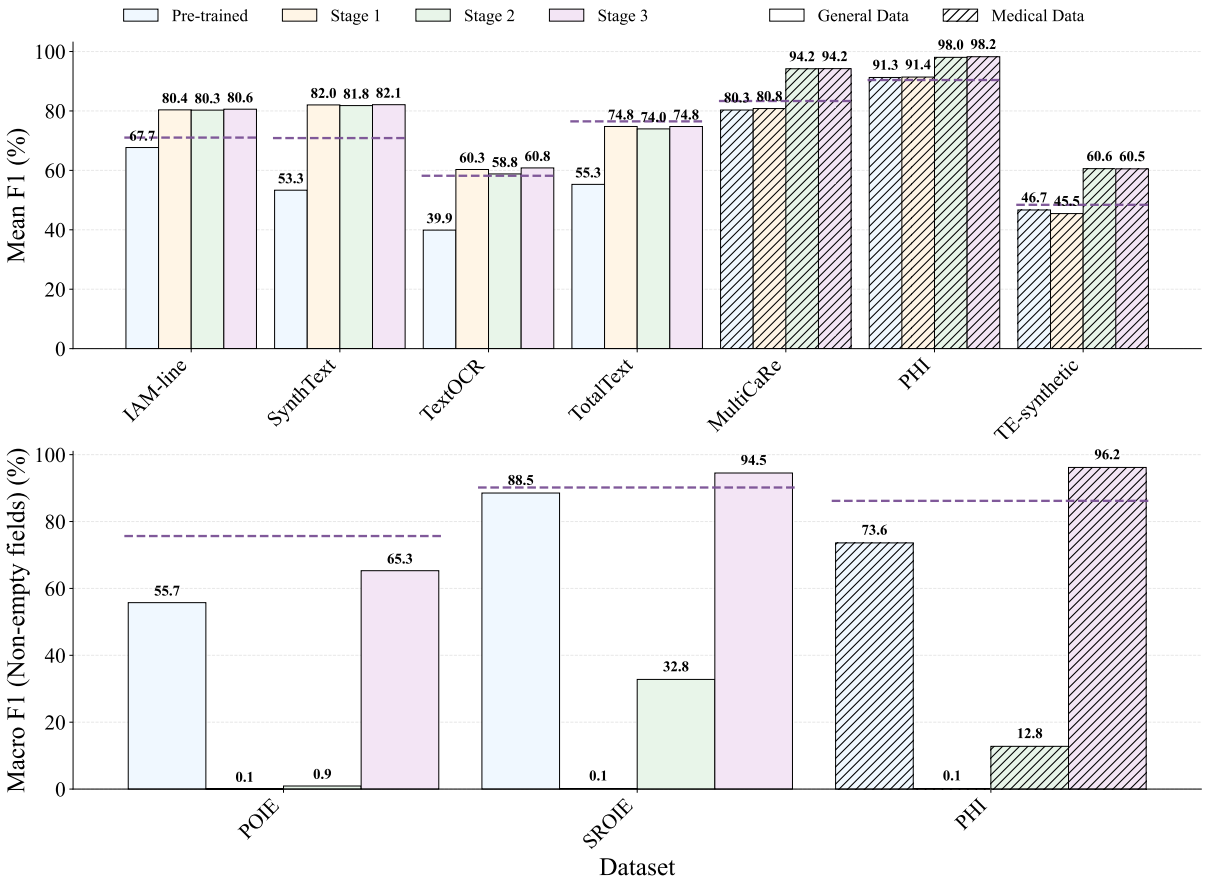

Figure 2: Performance dynamics of MedGemma 4B across fine-tuning stages. Top: mean F1 (%) for text extraction datasets. Bottom: macro F1 for non-empty fields (%) for key information extraction datasets. Bars are color-coded by stage; hatched patterns indicate medical datasets. Dashed lines indicate the Gemma 27B zero-shot baseline.

trained KIE baseline (e.g., 94.51% vs. 90.18% on SROIE, as shown in Figure 2).

This failure mode is practically significant for OCR systems that are expected to produce a variety of output structures—from plain transcription and multi-level JSON for structurally complex forms to LaTeX rendering for mathematical documents or domain-specific markup for document parsing. It calls for a careful composition of the fine-tuning corpus to integrate the output formats representative of the intended deployment scenarios. The low remediation cost implies that the bias could be mitigated with a limited amount of supervision.

4.4 Hallucination Reduction and Behavioral Stabilization

Fine-tuning produces qualitative behavioral changes critical for deployment reliability alongside accuracy gains. Pre-trained models exhibit three main failure types: repetitive over-generation triggered by complex image conditions or repeated textual content; generic fallback responses (e.g., "The image shows...", "I am unable to transcribe the text"); and content guessing when text is illegible rather than returning empty output.

We use predictions of at least twice the ground truth length as an operational indicator of uncontrolled repetition and over-generation, which is the dominant hallucination failure mode we observe. This proxy is intended to capture only extreme length outliers that inflate error metrics rather than content errors at a fine-grained level—this is detected by the standard evaluation metrics used in this work.

Fine-tuning reduces the proxy values significantly across most datasets, for example SynthText (222 $\rightarrow$ 53), TE-synthetic (182 $\rightarrow$ 4), MultiCaRe-TE (95 $\rightarrow$ 18), and PHI-TE (13 $\rightarrow$ 0). Fallback responses sharply decline—on MultiCaRe-TE, predictions on empty-ground-truth samples decrease from 818 to 5 after fine-tuning. The main exception is TextOCR, where the hallucination proxy increases (103 $\rightarrow$ 382) despite a reduction in trimmed CER. Qualitative inspection suggests that after fine-tuning, repetitive generation occurred more frequently for cases containing repeated strings

Dataset		Models									
		MedGemma 4B				Gemma 4B				Gemma 27B	EasyOCR
		Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Baseline
(a) Text Extraction — mean F1 (%); higher is better											
General	TextOCR	39.89	60.29	58.80	**60.82**	49.45	58.86	57.85	59.15	58.18	25.98
	TotalText	55.29	74.77	73.96	74.76	69.87	68.02	66.78	67.25	**76.49**	32.64
	IAM-line	67.69	80.37	80.28	**80.59**	66.42	73.46	73.85	73.40	71.03	10.93
	SynthText	53.33	82.03	81.81	**82.11**	65.58	78.89	79.00	79.89	70.87	37.80
Medical	TE-synthetic	46.67	45.46	**60.57**	60.51	42.69	35.99	52.59	52.67	48.41	22.32
	PHI-TE	91.27	91.40	98.04	**98.24**	87.32	85.38	97.34	97.47	90.41	68.58
	MultiCaRe-TE	80.31	80.80	94.19	**94.23**	65.72	76.85	93.45	93.46	83.35	39.04
(b) Key Information Extraction — macro F1 for non-empty fields (%); higher is better											
General	SROIE	88.52	0.15	32.79	**94.51**	85.26	0.17	16.44	92.18	90.18	—
	POIE	57.62	0.05	0.92	65.30	59.84	0.05	0.86	63.35	**75.66**	—
Medical	PHI-KIE	73.61	0.15	12.78	**96.18**	81.42	0.08	0.08	95.05	86.18	—

Table 2: Results on the curated OCR corpus across all models and fine-tuning stages. Bold: best per dataset row. EasyOCR is not evaluated on the KIE datasets (—), as it is not designed for structured extraction.

in ground truth, common for TextOCR. In such cases, the model repeatedly samples the same token instead of appropriately generating the end-of-sequence token.

The implications of errors depend on the downstream task. For PHI identification, repetitive but semantically relevant predictions may still allow a classifier to correctly determine the PHI category. Similarly, even if a prediction contains the correct text but is not represented in the expected format, it could be sufficient for the classifier. At the same time, meaningless hallucinations or guesses might be problematic, and we observe a notable reduction of such failures through fine-tuning.

4.5 Specialization–Generalization Trade-off on CC-OCR

On CC-OCR subsets resembling the fine-tuning corpus, Stage-3 MedGemma 4B recovers a substantial portion of the gap to Gemma 27B, reaching 93.1% of the large model's performance overall (Table 3). Fine-tuning benefits therefore extend beyond the curated corpus when document characteristics and output requirements remain similar.

On more divergent subsets, including multilingual OCR, document parsing, and tasks with unseen output structures, the fine-tuned model underperforms both Gemma 27B and its own pre-trained version. These results quantify the specialization cost and define a clear deployment boundary: compact fine-tuned models are effective in narrow, predictable OCR settings, but specialization reduces robustness to unseen scripts, document structures, and output regimes.

Dataset / Group	MedGemma 4B PT	MedGemma 4B FT (S3)	Gemma 27B PT	FT/27B (%)
Similar to training distribution (mean F1, %)				
CORD (Document)	76.9	80.2	91.5	87.5
FUNSD (Document)	72.5	71.0	81.4	87.2
IC15 (Scene)	51.8	55.8	60.7	92.0
InverseText (Scene)	59.7	71.7	72.3	99.1
LAION-OCR (UGC)	80.5	85.5	85.3	100.2
Average (5 datasets)	68.3	72.9	78.3	93.1
Diverging from training distribution				
Multi-scene OCR (F1)	41.7	23.7	64.4	36.8
Doc. parsing (edit sim.)	25.3	14.2	45.1	31.5
Multilingual OCR (F1)	57.9	55.5	73.1	75.9
KIE (micro F1)	27.2	21.2	43.9	48.3

Table 3: CC-OCR generalization evaluation. Rightmost column denotes the fine-tuned MedGemma's performance as a percentage of the 27B baseline.

4.6 Downstream PHI Identification Pipeline

In the PHI identification pipeline of Truong et al. (2025b), replacing the OCR component with Stage-3 MedGemma 4B improves recall to 99.35%, compared with 96.71% for EasyOCR, 96.84% for pre-trained MedGemma 4B, and 97.17% for Gemini 2.5 Flash when each is used as the OCR extractor (Table 4). False negatives drop from 49 to 10, while false positives also decrease from 30 to 24.

Model	FP	FN	Prec.	Rec.
MedGemma 4B FT (S3)	**24**	**10**	**98.47**	**99.35**
MedGemma 4B PT	30	49	98.04	96.84
EasyOCR	70	51	95.53	96.71
Gemini 2.5 Flash	43	44	97.25	97.17

Table 4: PHI identification pipeline results. The listed models replace the OCR extractor; Gemini 2.5 Flash is the fixed PHI classifier. FP/FN are absolute counts; Precision/Recall are percentages. Bold: best per metric.

Because missed PHI is the primary safety risk in clinical de-identification, this provides evidence that OCR gains translate into practically meaningful downstream improvements. This downstream improvement is consistent with large reductions in raw OCR errors on RadPHI-test (see Appendix C.3 for full per-model OCR metrics). Localizing the OCR component reduces the data-exposure surface of the pipeline, since only short extracted text strings are passed to the downstream classifier. Although it is a step toward privacy-preserving deployment, the pipeline is not fully local because the PHI classifier remains external.

5 Discussion

Our results show that large proprietary VLMs are not the only viable option for medical OCR. Within a controlled comparison, compact open-weight VLMs adapted with LoRA match or exceed a much larger zero-shot generalist on targeted text extraction and key information extraction tasks, while also improving a clinically relevant downstream PHI identification pipeline. In privacy-sensitive settings with relatively stable task requirements, compact specialized OCR components can provide strong performance without relying on a large external OCR model.

At the same time, the results define a clear deployment boundary. On the curated corpus, fine-tuned compact models perform strongly when document characteristics, domain, and output structure are well aligned with training. On CC-OCR, however, the same specialization reduces robustness on more heterogeneous inputs, particularly unseen scripts, structurally different document tasks, and unfamiliar output regimes. Compact fine-tuned models therefore appear most advantageous in OCR settings where the input distribution, language space, and output requirements are reasonably predictable, whereas large generalist models remain preferable when robustness across diverse formats and languages is a primary requirement.

The staged design further shows that OCR adaptation changes model behavior beyond transcription accuracy. It reduces severe over-generation, improves stability, and can alter output formatting behavior. This is practically important because OCR often serves as one component in a larger pipeline. For example, in the PHI identification pipeline, OCR improvements translated directly into fewer missed PHI instances, the key safety objective in de-identification.

Finally, our findings suggest that domain-aligned pre-training plays a role in the OCR model selection. Within the target domain it can provide a strong basis for zero-shot performance and compensate for scale differences, but at the cost of general breadth. Furthermore, it can be advantageous for OCR adaptation in the Gemma model family as suggested by the observed amplified PEFT gains, although this interpretation should remain cautious because the composition of the pre-training corpus is not fully disclosed.

Overall, the study supports compact OCR systems specialized for clearly defined deployment contexts. In privacy-sensitive domains such as healthcare, this provides a pathway toward reducing dependence on external proprietary APIs while maintaining competitive OCR performance.

6 Conclusion

We investigate whether compact open-weight VLMs can be adapted into strong medical OCR systems through parameter-efficient fine-tuning with LoRA, and whether such systems can reduce reliance on large proprietary models in privacy-sensitive clinical deployments. The results are broadly positive within well-aligned OCR settings, though they reveal a specialization-generalization trade-off that outlines the boundary of applicability. Fine-tuned MedGemma 4B surpasses the untuned 27B generalist on most targeted benchmarks and outperforms Gemini 2.5 Flash in a downstream PHI identification pipeline, reducing missed PHI instances by 80%. Domain-aligned pre-training provides a consistent advantage both before and during fine-tuning. Output format overfitting—a side effect of single-format fine-tuning—is mitigated with minimal multi-task supervision. These findings support a shift toward compact, specialized, and locally deployable VLMs as OCR components for defined clinical workflows.

Limitations

Single specialization strategy. We focus exclusively on supervised fine-tuning as a mechanism for OCR specialization. Future research should examine how it compares to the gains and costs of alternative approaches such as few-shot prompting.

Single model family. All experiments use models from the Gemma 3 family, with MedGemma 4B as the sole domain-specialized variant. Whether the observed patterns, such as amplified PEFT gains from domain-aligned pre-training hold for other architectures and medical VLMs remains to be established. Comparisons against small VLMs specifically pre-trained for OCR were outside the scope of this work; it is an open question whether OCR-task pre-training or domain pre-training provides a stronger initialization for medical OCR fine-tuning.

Unknown pre-training corpus. The composition of MedGemma's biomedical pre-training data is not disclosed, which prevents interpretation of the domain alignment effects we observe. The hypothesis that biomedical pre-training instills a transferable OCR prior is empirically supported but cannot be causally confirmed without access to pre-training details or controlled ablations.

Visual encoder constraint. The SigLIP encoder used by the Gemma 3 family resizes all inputs to a fixed 896×896 resolution, which limits recognition accuracy on low-resolution image crops. Performance on datasets such as TE-synthetic remained comparatively low despite fine-tuning, possibly reflecting an architectural ceiling that PEFT cannot overcome. Future work using VLMs with native-resolution or aspect-ratio-preserving encoders may achieve better results on such inputs.

Synthetic medical data. The scarcity of real annotated medical OCR data required the use of synthetic datasets for both training and evaluation. While the generation pipelines were designed for realism and diversity, real clinical data would be preferable both for fine-tuning and benchmarking.

Ethical Considerations

This study is a retrospective analysis of open-access OCR datasets together with synthetically generated medical OCR data. No new participant recruitment or direct data collection was performed by the authors. We used the underlying datasets in accordance with their stated access conditions and license terms. The work is intended for research and evaluation only, and any downstream deployment should be validated on institution-specific data under applicable governance policies.

References

Mahaman Sanoussi Yahaya Alassan, Jessica López Espejel, Merieme Bouhandi, Walid Dahhane, and El Hassane Ettifouri. 2024. Comparison of open-source and proprietary LLMs for machine reading comprehension: A practical analysis for industrial applications. *Preprint*, arXiv:2406.13713.

Edouard Belval. 2016. TextRecognitionDataGenerator: a synthetic data generator for text recognition. https://github.com/Belval/TextRecognitionDataGenerator. GitHub repository.

Jiawei Chen, Yue Jiang, Dingkang Yang, Mingcheng Li, Jinjie Wei, Ziyun Qian, and Lihua Zhang. 2024. Can LLMs' tuning methods work in medical multimodal domain? In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part V*, pages 112–122, Berlin, Heidelberg. Springer-Verlag.

Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2025. Ocean-OCR: Towards general OCR application via a vision-language model. *Preprint*, arXiv:2501.15558.

Chee Kheng Chng and Chee Seng Chan. 2017. Total-text: A comprehensive dataset for scene text detection and recognition. *Preprint*, arXiv:1710.10400.

Daniele Faraglia and Other Contributors. Faker.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. Preprint, arXiv:2503.19786.

Gavin Greif, Niclas Griesshaber, and Robin Greif. 2025. Multimodal LLMs for OCR, OCR post-correction, and named entity recognition in historical documents. *Preprint*, arXiv:2504.00414.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Unnati Gupta, Ayush Kumar, Arti Gupta, Gaurav Raj, and Arun Prakash Agrawal. 2024. Advances in handwritten character recognition: A comparison of OCR and large language model-based approaches. In *Proc. Int. Conf. Emerging Technologies and Innovation for Sustainability (EmergIN)*, pages 192–198, Greater Noida, India. IEEE.

Vincent Hanke, Tom Blanchard, Franziska Boenisch, Iyiola Olatunji, Michael Backes, and Adam Dziedzic. 2024. Open LLMs are necessary for current private adaptations and outperform their closed alternatives. *Advances in Neural Information Processing Systems*, 37:1220–1250.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR 2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*.

Jaided AI. 2020. EasyOCR: a ready-to-use OCR with 80+ supported languages. `https://github.com/JaidedAI/EasyOCR`. GitHub repository.

Pooja Jain, Kavita Taneja, and Harmunish Taneja. 2021. Which OCR toolset is good and why: A comparative study. *Kuwait Journal of Science*, 48(2).

Daniel P. Jeong, Pranav Mani, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. 2025. The limited impact of medical adaptation of large language and vision-language models. *Preprint*, arXiv:2411.08870.

Yichao Jin, Yushuo Wang, Qishuai Zhong, Kent Chiu Jin-Chun, Kenneth Ke Zhu, and Donald MacDonald. 2025. Multi-stage field extraction of financial documents with OCR and compact vision-language models. *Preprint*, arXiv:2510.23066.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. *Preprint*, arXiv:2305.07498.

Anand Kumar, Pardeep Singh, and Kusum Lata. 2023. Comparative study of different optical character recognition models on handwritten and printed medical reports. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 581–586.

Ernestine Lerisha J, Santhosh M. R. S, Perumalraja Rengaraju, Chung-Horng Lung, and Yang Cao. 2025. Automation of medical insurance post-claims using OCR and RAG models. In *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 2026–2031.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. 2025. Visual large language models for generalized and specialized applications. *Preprint*, arXiv:2501.02765.

Yunyi Liu, Yingshu Li, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2024. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. *Meta-Radiology*, 2(4):100099.

Urs-Viktor Marti and Horst Bunke. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Sankalp Nagaonkar, Augustya Sharma, Ashish Choithani, and Ashutosh Trivedi. 2025. Benchmarking vision-language models on optical character recognition in dynamic video environments. *Preprint*, arXiv:2502.06445.

Mauro Nievas Offidani. 2024. MultiCaRe: An open-source clinical case dataset for medical image classification and multimodal AI applications (2.0).

Dhavalkumar Patel. 2025. Comparing traditional OCR with generative AI-assisted OCR: Advancements and applications. *International Journal of Science and Research (IJSR)*, pages 347–351.

Mahesh Ponnuru, P. Sridevi Ponmalar, A. Likhitha, B. Tanu Sree, and G. Guna Chaitanya. 2024. Image-based extraction of prescription information using OCR-Tesseract. *Procedia Computer Science*, 235:1077–1086. International Conference on Machine Learning and Data Engineering (ICMLDE 2023).

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, and others. 2025. MedGemma technical report. *Preprint*, arXiv:2507.05201.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. *Preprint*, arXiv:2105.05486.

Iknoor Singh, Miguel Colom, and Kalina Bontcheva. 2025. A comparative analysis of OCR models on diverse datasets: Insights from memes and Hiertext dataset. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1253–1263.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2025. Are small language models ready to compete with large language models for practical applications? *Preprint*, arXiv:2406.11402.

Salvatore Tabbone and Matthieu Pelingre. 2025. Benchmarking OCR tools for historical postcards: A dataset and evaluation. In *Proceedings of the 7th International Workshop on analySis, Understanding and proMotion of heritAge Contents*, SUMAC '25, pages 61–68, New York, NY, USA. Association for Computing Machinery.

Alexandru Cristian Telea. 2004. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9:23–34.

Tuan Truong, Ivo M. Baltruschat, Mark Klemens, Grit Werner, and Matthias Lenga. 2025a. Exploring AI-based system design for pixel-level protected health information detection in medical images. *Journal of Imaging Informatics in Medicine*, 39(2):1796–1815.

Tuan Truong and Matthias Lenga. 2025. Pixel-level protected health information (PHI) dataset.

Tuan Truong, Guillermo Jimenez Perez, Pedro Osorio, and Matthias Lenga. 2025b. Towards selection of large multimodal models for burned-in protected health information detection in medical images. *Preprint*, arXiv:2511.02014.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Zhibo Yang, Jun Tang, Zhaohai Li, and others. 2024. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Liqiao Xia, Jeremy Yeung, Daochen Zha, Dongming Cai, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. 2025. Large language models for disease diagnosis: A scoping review. *Preprint*, arXiv:2409.00097.

A Data Corpus and Pre-processing Details

A.1 Corpus Contents

Figure 3 illustrates the contents of the curated OCR corpus.

A.2 Corpus Composition

Table 5 provides the complete corpus overview and the corresponding train/validation/test partitions. The rules for partitioning are as follows: if original dataset splits are unavailable, samples are randomly partitioned using a 70/10/20 ratio. For datasets containing only train and validation splits, the validation set is treated as test data and 10% of the training data is reallocated for validation. For datasets with predefined train/test splits, 10% of the training data is assigned for validation. For the datasets with the three predefined splits, the original structure is preserved. TextOCR is an exception: because it lacks test labels, the original validation split serves as test data, and a new validation split is extracted from the training set. To avoid data leakage, samples derived from the same base images (SynthText) or the same patients (MultiCaRe-TE) are assigned to the same split.

A.3 Synthetic Data Generation

TE-synthetic. Text strings are generated using the `Faker` library (Faraglia and Other Contributors) across categories including names, emails, dates, and addresses, to simulate realistic PHI-like content. These are rendered onto randomly cropped radiological image backgrounds using `TextRecognitionDataGenerator` (Belval, 2016) with varied fonts, sizes, and distortions, reflecting the prevalence of small, low-contrast imprints in radiological scans.

MultiCaRe-TE. Sentences from image captions in the MultiCaRe corpus (Nievas Offidani, 2024) are overlaid onto their associated images using a custom pipeline. Placement uses an adaptive strategy, progressively relaxing margin and font-size constraints to find valid non-overlapping regions.

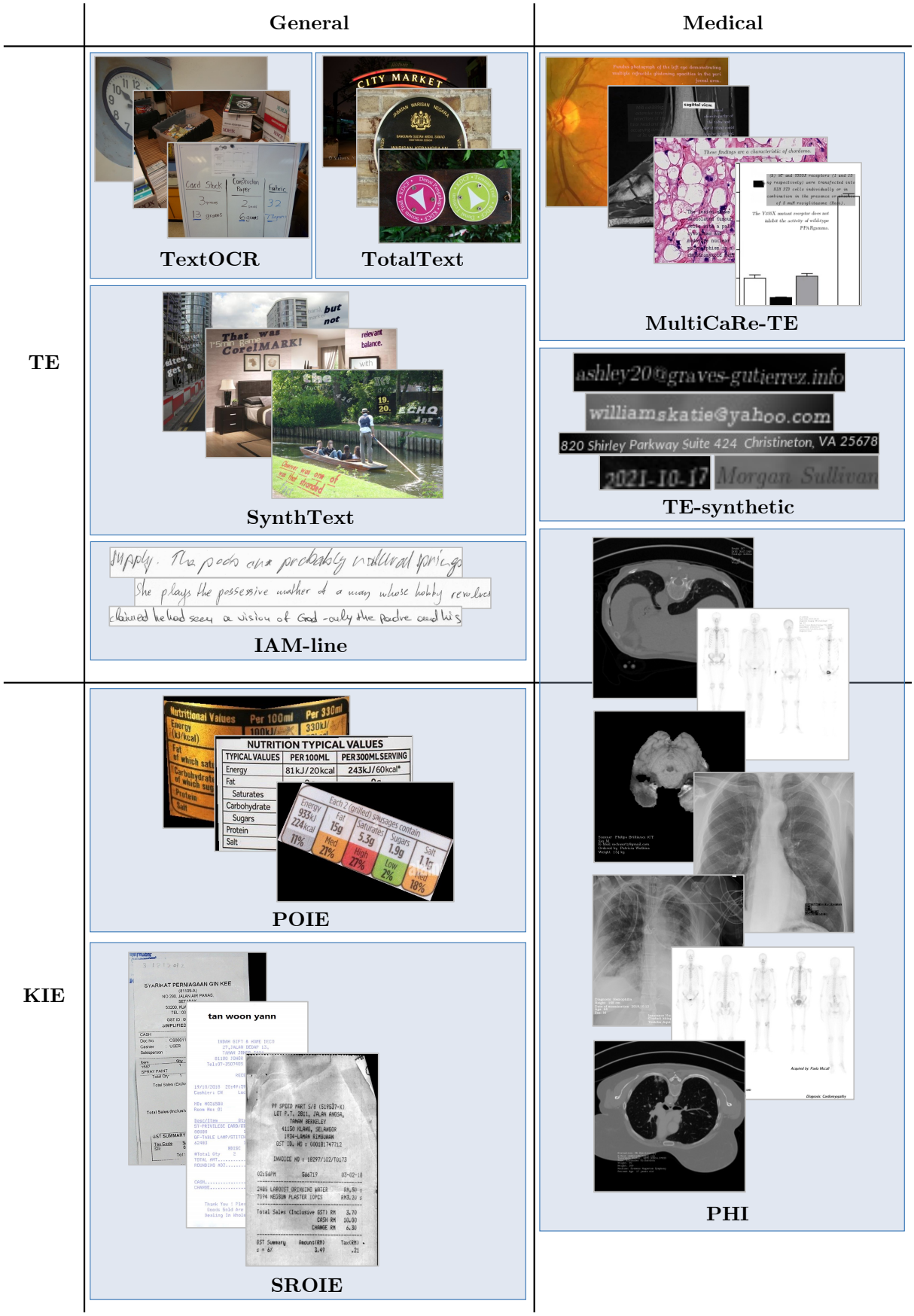

Figure 3: Examples of samples from the curated OCR corpus across domains (general and medical) and tasks (text extraction and key information extraction).

Dataset	Task	Domain	Data Characteristics	Train	Val	Test	Total
IAM-line (Marti and Bunke, 2002)	TE	General	Handwritten text lines	6,482	976	2,915	10,373
TotalText (Chng and Chan, 2017)	TE	General	Curved and multi-oriented scene text	1,129	126	300	1,555
TextOCR (Singh et al., 2021)	TE	General	General scene text	19,411	2,161	3,085	24,657
SynthText (Gupta et al., 2016)	TE	General	Synthetic rendered scene text	62,337	8,905	17,811	89,053
MultiCaRe-TE†	TE	Medical	Cross-specialty medical images with overlaid captions	70,959	9,755	20,524	101,238
TE-synthetic	TE	Medical	Radiological crops with short PHI-like text imprints	13,500	1,500	6,000	21,000
PHI-TE (Truong et al., 2025a)	TE	Medical	Radiology images with simulated text imprints	4,900	700	1,400	7,000
SROIE (Huang et al., 2019)	KIE	General	Scanned receipts, 4 fixed fields	563	63	347	973
POIE (Kuang et al., 2023)	KIE	General	Nutritional product labels, up to 21 fields	2,025	225	750	3,000
PHI-KIE (Truong et al., 2025a)	KIE	Medical	Radiology images with PHI key-value pairs, 17 fields	4,900	700	1,400	7,000
Total				**186,206**	**25,111**	**54,532**	**265,849**

† Generated from the MultiCaRe corpus (Nievas Offidani, 2024); see Appendix A.3.

Table 5: Overview of the curated OCR corpus and corresponding splits.

Font size is scaled proportionally to image dimensions and clamped to 6–18pt. Text color is selected based on background luminance contrast (minimum 70 intensity units from local mean for homogeneous regions; K-means dominant color analysis for heterogeneous regions). Background boxes are applied with a base probability of 30% (increased for thin or italic fonts). Boxes use randomized padding and grayscale colors; opacity is full in 65% of cases, otherwise sampled uniformly from $[80, 220]$. For transparent boxes, perceived color for text color selection is computed as $C_{\text{final}} = (1 - \alpha/255) \cdot C_{\text{patch}} + (\alpha/255) \cdot C_{\text{box}}$. Rotation is applied with 50% probability using small angles sampled from $\mathcal{N}(0, 1.5)$.

A.4 Image Cleaning and Inpainting

Several datasets from the curated corpus contain visible but non-annotated text regions that can bias evaluation by penalizing correct predictions absent from the ground truth annotations. Such regions are identified and removed through image inpainting.

For TextOCR and TotalText, non-annotated text regions are explicitly marked in the annotations (“.” and “#”, respectively). For MultiCaRe, no such annotations are available. Consequently, non-annotated textual regions are detected automatically using PaddleOCR and YOLO-based detectors, after which the resulting bounding boxes are merged for maximal coverage.

The identified regions are removed using OpenCV’s `INPAINT_TELEA` algorithm (Telea, 2004). Figure 4 illustrates the procedure.

The proportion of images requiring inpainting is approximately 63% for MultiCaRe, 73% for TextOCR, and 42% for TotalText. Although some text might remain undetected, this step should substantially improve evaluation reliability.

Detected bounding boxes (PaddleOCR: green, YOLO: blue) | Generated mask | Final inpainted image

Figure 4: Illustration of the image inpainting pipeline used to remove non-annotated textual regions.

A.5 Standardized Annotation Schema

To enable consistent processing across datasets, we design a unified JSON schema that standardizes transcription fields, bounding boxes, and metadata (Listing 1). Many fields are retained from the original dataset annotations for completeness even when not used in our framework.

```
{
  "image_id": "dataset-name_original-id"
      ,
  "file_name": "original-image-filename"
      ,
  "image_path": "/path/to/image",
  "source_dataset": "string",
  "original_split": "train | val | test
      | None",
  "width": "int",
  "height": "int",
  "domain": "scene_text | handwritten |
      medical | ...",
```

```
  "modality": "CT | MRI | XRay |
      Ultrasound | other | None",
  "patient_id": "string | None",
  "license": "string | None",
  "annotations": [
    {
      "annotation_id": "string",
      "text_content": "string",
      "semantic_tag": "text_region |
          timestamp | ...",
      "bbox": ["x_min", "y_min", "x_max"
          , "y_max"],
      "polygon": [["x1", "y1"], ..., ["
          xN", "yN"]],
      "text_orientation":
        "horizontal | vertical | curved
            | None",
      "metadata": { "phi": "0 | 1" }
    }
  ],
  "key_value_pairs": {
    "key_name": "extracted_value"
  }
}
```

Listing 1: Unified annotation schema used across the curated OCR corpus.

B Implementation Details

B.1 Prompt Templates

Text Extraction Prompt. A single unified prompt is used across all TE datasets:

```
Task: Transcribe the text in this image.
Follow these rules:
1.  Look carefully and identify all
visible text.
2.  Transcribe exactly as it appears,
preserving capitalization, spacing, and
punctuation.
3. Do NOT correct spelling or grammar.
4.  Do NOT add commentary, description,
or formatting.
5.  Do NOT guess or fill in text not
clearly visible. If part of the text is
obscured, leave it blank.
6. Respond only with the extracted text.
If there is no visible text, return an
empty string.
```

Key Information Extraction Prompt. Dataset-specific prompts are constructed using the following template, instantiated with the relevant field list and JSON schema for each dataset:

```
Task:  Extract the following [context]
fields from this image.
Possible fields (may or may not appear
in each sample): [List of Fields]

Rules:
1. Use exactly the text from the image.
2.   If a field is not present or
unreadable, output an empty string.
3.    Respond ONLY with valid JSON
containing these [N] keys.

Output format: [JSON schema]
Now extract from this image:
```

The specific KIE configurations are as follows:

SROIE

The four target receipt fields are `company`, `date`, `address`, and `total`. For the date field, an additional rule is included: "For the date field, keep the format as it appears in the image. The date field is designed to display only the date, without including the specific time."

PHI-KIE

The schema includes 17 medical fields: `hospital`, `identifier`, `marker`, `phone_nr`, `diagnosis`, `height`, `gender`, `patient_name`, `scanner`, `date_of_birth`, `imaging_personnel`, `weight`, `address`, `exam_date`, `age`, `email`, and `examination_type`.

POIE

The prompt instructs to extract nutritional values listed in a product label. The schema specifies 21 keys: `SS` (Serving Size) plus four variants (`PS`, `P1`, `D`, `PP`) for each of the five nutritional components: `CE` (Calories/Energy), `TF` (Total Fat), `SO` (Sodium), `CAR` (Carbohydrates), and `PRO` (Protein).

The original ground truth annotations for the POIE dataset use abbreviated field names (e.g., "CE-PP", "TF-D") without explicitly defining them in the dataset documentation. To construct meaningful prompts, we derive definitions by manually cross-referencing the ground truth values with the original product images.

During this process, we identified potential discrepancies in the way these abbreviations have been interpreted in prior benchmarks. Namely, the suffix "PP", interpreted as "per package" by Yang et al. (2024), visually corresponds to "% adult RI per 100g/ml", aligning with OCRBench (Fu et al., 2025). Similarly, the suffix "D", interpreted as "daily" by CC-OCR, appears to specifically denote "% daily value/intake *per serving*". These interpretations appear most consistent with the data; however, without official documentation, they remain inferential.

B.2 Hyperparameter Tuning

Learning rate. We tune the learning rate with LoRA rank fixed at $r = 8$, $\alpha = 16$ using

MedGemma 4B on TextOCR—the most challenging model-dataset combination at baseline. Four values are evaluated: 1×10^{-4}, 5×10^{-5}, 3×10^{-5}, 1×10^{-5}. The highest rate causes rapid overfitting; the lowest causes underfitting. The value 3×10^{-5} provides optimal validation performance and is kept fixed for all experiments.

LoRA rank and scaling factor. A grid search is first conducted on TextOCR and subsequently on the full Stage-1 validation corpus (TextOCR, TotalText, IAM-line, SynthText) for both Gemma 4B and MedGemma 4B. Performance is aggregated using average Z-scores across datasets. Tables 6 and 7 report the results. The $\alpha = r/2$ configuration consistently underperforms; $\alpha = 2r$ is most commonly best. Higher ranks improve performance with diminishing returns. The configuration $r = 64$, $\alpha = 128$ is the strongest within memory constraints and is used for all fine-tuning stages.

	$\alpha = 4$	$\alpha = 8$	$\alpha = 16$	$\alpha = 32$
$r = 8$	58.53	59.33	59.67	—
$r = 16$	—	59.74	60.38	60.10

Table 6: LoRA hyperparameter search on TextOCR validation set (MedGemma 4B). Mean F1 (%); higher is better.

MedGemma 4B

	$\alpha = r/2$	$\alpha = r$	$\alpha = 2r$
$r = 16$	—	-0.364	-0.801
$r = 32$	-0.298	-0.472	0.933
$r = 64$	-0.781	0.185	1.596

Gemma 4B

	$\alpha = r/2$	$\alpha = r$	$\alpha = 2r$
$r = 16$	—	-0.181	-1.161
$r = 32$	-0.443	-0.479	0.478
$r = 64$	—	0.250	1.538

Table 7: LoRA hyperparameter search on the Stage-1 validation corpus Reporting average Z-scores of normalized mean F1 across datasets; higher is better.

B.3 Hyperparameter Configuration

Table 8 summarizes the final training configuration used for fine-tuning Gemma 4B and MedGemma 4B throughout all three stages.

Parameter	**Value**
Training Configuration (SFTTrainer)	
Optimizer	AdamW
Learning rate	3×10^{-5}
LR scheduler	Cosine annealing
Warmup ratio	0.05
Weight decay	0.01
Per-device batch size	4
Gradient accumulation	8 steps
Effective batch size	32
Max gradient norm	1.0
Precision	bfloat16
Eval/save strategy	Every 900 steps
Checkpoint selection	Min. validation loss
Loss function	Cross-entropy
LoRA Configuration	
Task type	`CAUSAL_LM`
Rank (r)	64
Alpha (α)	128
Dropout	0.1
Bias	None
Target modules	`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`
Adapter init	Fresh at each stage

Table 8: Hyperparameters for all fine-tuning stages.

C Additional Results

C.1 Secondary TE and KIE Metrics

Table 9 reports CER and WER for the TE task. Table 10 reports ANLS and binary F1 for the KIE task (EasyOCR is omitted as it cannot perform key information extraction).

C.2 Performance Dynamics Across Fine-Tuning Stages

Table 11 reports the percentage improvement at each fine-tuning stage relative to the pre-trained baseline, and the stage-to-stage change in percentage points, for Gemma 4B and MedGemma 4B.

C.3 Isolated OCR Capability in the PHI Pipeline

Table 12 reports OCR capability evaluated in isolation on text crops from RadPHI-test, comparing MedGemma 4B against selected models from Truong et al. (2025b) evaluated under the same metric protocol.

Trimmed Mean (10%) CER

TE datasets		Models									
		MedGemma 4B				Gemma 4B				Gemma 27B	EasyOCR
		Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Baseline
General	TextOCR	0.611	0.419	0.549	**0.377**	0.497	0.361	0.390	0.355	0.381	0.651
	TotalText	0.363	0.147	0.155	0.153	0.204	0.224	0.223	0.205	**0.135**	0.546
	IAM-line	0.317	0.184	0.183	**0.180**	0.341	0.258	0.250	0.254	0.290	0.735
	SynthText	0.495	0.176	0.177	**0.176**	0.374	0.214	0.213	0.197	0.298	0.558
Medical	TE-synthetic	0.304	0.322	**0.179**	0.180	0.399	0.480	0.275	0.279	0.344	0.629
	PHI-TE	0.038	0.034	0.002	**0.001**	0.075	0.067	0.003	0.003	0.027	0.215
	MultiCaRe-TE	0.150	0.129	0.023	**0.023**	0.324	0.171	0.028	0.028	0.107	0.519

Trimmed Mean (10%) WER

TE datasets		Models									
		MedGemma 4B				Gemma 4B				Gemma 27B	EasyOCR
		Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Baseline
General	TextOCR	0.730	0.602	0.757	0.547	0.631	0.546	0.593	0.536	**0.525**	0.858
	TotalText	0.531	0.292	0.302	0.296	0.350	0.406	0.421	0.406	**0.271**	0.821
	IAM-line	0.428	0.233	0.233	**0.228**	0.447	0.324	0.317	0.323	0.377	1.048
	SynthText	0.590	0.212	0.214	**0.211**	0.438	0.261	0.263	0.241	0.362	0.761
Medical	TE-synthetic	0.600	0.623	**0.408**	0.410	0.680	0.846	0.513	0.515	0.600	0.876
	PHI-TE	0.074	0.072	0.006	**0.005**	0.105	0.121	0.010	0.010	0.053	0.359
	MultiCaRe-TE	0.187	0.172	0.034	**0.034**	0.362	0.235	0.040	0.040	0.169	0.697

Table 9: TE evaluation results under a 10% outlier rate protocol. Top: Character Error Rate (CER); Bottom: Word Error Rate (WER). Lower is better. Bold: best result per dataset.

Macro ANLS (%)

KIE datasets		Models								
		MedGemma 4B				Gemma 4B				Gemma 27B
		Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained
General	SROIE	88.28	0.11	27.09	**95.05**	85.10	0.19	13.81	93.86	88.16
	POIE	61.11	0.06	0.91	67.85	60.15	0.06	0.83	66.04	**75.30**
Medical	PHI-KIE	74.57	0.11	12.68	**97.37**	83.33	0.02	0.03	96.71	87.93

Macro Binary F1 (%)

KIE datasets		Models								
		MedGemma 4B				Gemma 4B				Gemma 27B
		Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained	Stage 1	Stage 2	Stage 3	Pre-trained
General	SROIE	99.96	0.72	73.16	**100.00**	99.96	0.72	46.67	99.93	99.96
	POIE	61.38	9.68	32.36	79.90	66.11	9.68	27.92	**83.21**	71.43
Medical	PHI-KIE	84.58	59.56	20.22	98.81	86.66	82.82	76.83	**98.95**	92.64

Table 10: KIE evaluation results. Top: Macro ANLS for non-empty fields (%); Bottom: Macro binary F1 (%). Higher is better. Bold: best result per dataset.

	MedGemma 4B					Gemma 4B				
	vs. PT (%)			Stage-to-stage (pp)		vs. PT (%)			Stage-to-stage (pp)	
Dataset	S1	S2	S3	S1→S2	S2→S3	S1	S2	S3	S1→S2	S2→S3
Text Extraction (mean F1)										
TextOCR	51.1	47.4	52.5	−1.5	2.0	19.0	17.0	19.6	−1.0	1.3
TotalText	35.2	33.8	35.2	−0.8	0.8	−2.6	−4.4	−3.7	−1.2	0.5
SynthText	53.8	53.4	54.0	−0.2	0.3	20.3	20.5	21.8	0.1	0.9
IAM-line	18.7	18.6	19.1	−0.1	0.3	10.6	11.2	10.5	0.4	−0.4
PHI-TE	0.1	7.4	7.6	6.6	0.2	−2.2	11.5	11.6	12.0	0.1
MultiCaRe-TE	0.6	17.3	17.3	13.4	0.0	16.9	42.2	42.2	16.6	0.0
TE-synthetic	−2.6	29.8	29.7	15.1	−0.1	−15.7	23.2	23.4	16.6	0.1
Key Information Extraction (macro F1)										
SROIE	−99.8	−63.0	6.8	32.6	61.7	−99.8	−80.7	8.1	16.3	75.7
PHI-KIE	−99.8	−82.6	30.7	12.6	83.4	−99.9	−99.9	16.7	0.0	95.0
POIE	−99.9	−98.4	17.1	0.9	64.4	−99.9	−98.6	5.9	0.8	62.5

Table 11: Performance dynamics across fine-tuning stages (S1–S3). "vs. PT" denotes percentage improvement relative to the pre-trained baseline; "Stage-to-stage" denotes the absolute change in percentage points (pp) from the previous stage. TE results are based on mean F1 (%); KIE results on macro F1 for non-empty fields (%).

Metric	MedGemma 4B		InternVL2.5		Qwen2.5		GPT-4o	EasyOCR
	Stage 3	Pre-trained	4B	8B	3B	7B		
Character Error Rate (CER)								
Mean	0.0874	2.4686	0.6563	0.1165	0.0783	0.0415	0.0299	0.1179
Trimmed (10%)	0.0210	0.0619	0.0038	0.0030	0.0258	0.0018	0.0004	0.0610
Word Error Rate (WER)								
Mean	0.1326	1.1675	0.3172	0.0954	0.1353	0.0671	0.0471	0.2343
Trimmed (10%)	0.0564	0.1204	0.0143	0.0133	0.0692	0.0084	0.0015	0.1629
F1 Score								
Mean (%)	89.56	83.79	92.48	92.81	88.73	94.05	95.82	80.61

Table 12: Isolated OCR capability on RadPHI-test crops (Truong and Lenga, 2025). Trimmed CER/WER use 10% outlier rate. Comparative baseline numbers are compiled from Truong et al. (2025b), evaluated under an identical metric protocol.