Building Character(s): Synthetic Data and In-Context Learning Strategies for Few-Shot Ancient Chinese Recognition

Denise Atzori, Marie Bizais-Lillig, Mathias Garnier, Maxime Létoffé, Charles Planque, Tianjie Yin, Chahan Vidal-Gorène
École nationale des chartes – PSL
65 rue de Richelieu, 75002 Paris, France
surname.name@chartes.psl.eu

Abstract

Ancient Chinese character recognition remains challenging due to severe character imbalance, graphic variants, peculiar layout, degraded printing, and limited annotated data. This paper presents our system for EvaHan 2026, combining synthetic data generation and in-context learning (ICL) across three tasks: line-level text recognition (printed and handwritten) and page layout detection. We introduce *UltraGlyph*, a synthetic data pipeline recombining glyphs from real data with font-generated characters to improve rare-character coverage, producing 234,528 line images for foundation-model pretraining. We benchmark CRNN, transformer-based OCR, and a suite of vision–language models under a variant-aware ICL framework. On printed text, dedicated OCR systems and top VLMs reach comparable comprehensive scores with around 97% of accuracy; on cursive handwriting, performance drops significantly and is bounded above by 95%, with the best result achieved by Qwen2.5-VL-72B in zero-shot. For layout analysis, YOLO12s achieves the best score with a mAP50 of 75%.

Keywords: Historical Chinese OCR, Synthetic Data, VLM, Few-Shot Learning, In-Context Learning

1. Introduction

1.1. Text and Layout Recognition of Chinese historical Documents

Over the past decade, Chinese OCR and HTR research has shifted from modular pipelines toward end-to-end architectures. CRNN (Shi et al., 2015) established the paradigm of convolutional feature extraction with CTC decoding, followed by Transformer-based approaches such as `TrOCR` (Li et al., 2021) and Donut (Kim et al., 2022), which reformulated OCR as image-to-text generation. Frameworks like `LayoutLM` (Xu et al., 2020) and`DiT` (Li et al., 2022) further introduced layout-sensitive representations.

These advances have converged toward large-scale vision–language models (VLMs) as the dominant paradigm. `Qwen2-VL` (Wang et al., 2024) and `Qwen2.5-VL` (Bai et al., 2025) introduced dynamic-resolution ViT processing and M-RoPE, enabling robust Chinese OCR without task-specific redesign. A notable trend is the emergence of compact VLMs rivaling much larger models: `PaddleOCR-VL` (Cui et al., 2025), at only 0.9B parameters, achieves state-of-the-art parsing across 109 languages, while `CHURRO` (Semnani et al., 2025), a 3B-parameter model fine-tuned on 99,491 historical pages across 46 language clusters, demonstrates broad generalization across historical scripts, though historical Chinese remains its most challenging subset. `Xunzi-MLLM` (Zhu et al., 2025), fine-tuned on the *Siku quanshu* with a mixed ancient–modern corpus, offers a more targeted approach for classical Chinese. Combination of in-context learning (Brown et al., 2020) and retrieval-augmented demonstration selection (Gupta et al., 2023; Luo et al., 2024) has proven effective for low-resource script (Vidal-Gorène et al., 2026). LoRA-based fine-tuning (Hu et al., 2021) further enables adaptation under annotation constraints, with PEFT on `Qwen2.5-VL` showing strong synthetic-to-real transfer (Chung and Choi, 2025).

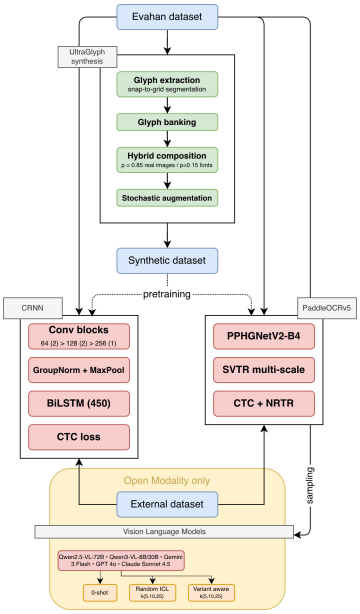


Figure 1: Overview of the pipeline (Tasks A and C)

Historical Chinese HTR nonetheless poses chal-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 339–352.

lenges that general-purpose VLMs do not resolve: extreme character-class imbalance across tens of thousands of sinograms and orthographic diversity due to graphic variants. Dedicated benchmarks —AncientDoc (Yu et al., 2025), MCS-Bench (Liu et al., 2025), and BABMLLM (Wang and Zhu, 2025), consistently confirm this gap: even the top-performing model on MCS-Bench (`InternVL2.5-78B`) achieves an average score below 50 across classical Chinese tasks. Synthetic data generation has emerged as a standard mitigation strategy, from CycleGAN-based style transfer (Chang et al., 2018) to large-scale datasets such as MegaHan97K (Zhang et al., 2025). A GAN-augmented CRNN with iterative rare-character targeting has achieved over 98% accuracy on xylographic editions (Bizais-Lillig et al., 2024), confirming that data quality and coverage remain the dominant factors regardless of architectural choice, and that reading order remains the principal challenge in complex layouts.

1.2. EvaHAN 2026 OCR Tasks

EvaHAN 2026 comprises three tasks, each available in closed and open modalities: (1) text recognition of printed xylographic lines (Dataset A), (2) page layout analysis via object detection (Dataset B), and (3) handwritten text recognition (Dataset C). Each dataset includes 5,000 training and 200 test images, targeting core challenges of historical Chinese document digitization: complex layouts, large glyph inventories, script variation, and degraded handwriting, with attention to vertical text and traditional, simplified, and variant character forms. In the closed modality, teams are restricted to the official training data and either `Qwen2.5-VL-7B-Instruct`, `Xunzi_Qwen2_VL_7B_Instruct`, or traditional ML models; external corpora and other LLMs are prohibited.

2. Datasets

2.1. Competition dataset

2.1.1. Dataset A

Training After minor corrections during the challenge, Dataset A comprises 4,997 low-quality line images sampled from standard xylographic editions (religion, philosophy, poetry, classical literature) plus paratextual material. Key challenges include: (1) **style diversity**: alongside standard prints, many lines contain idiosyncratic glyphs with variable stroke thickness, occasional blur, and a few color pages; (2) **imprecise cropping**: most images are line crops rotated 90° counterclockwise, leading to truncated strokes, bleed-in from neighboring characters, and visible grid lines resembling the character *yi* (one); (3) **orientation outliers**: a small subset is not rotated, requiring an orientation classifier (Section 3.1); (4) **multi-line artifacts**: a few samples contain two parallel lines or short horizontal sequences (while a pipeline using Vidal-Gorène et al. (2021) was developed, it was subsequently deprecated following the finalization of the test set specifications); (5) **glyph variants**: the usual historical-variant forms; (6) **blanks and punctuation** to be removed. The line-crop format provides limited context, likely affecting both human reading and OCR, though its impact remains to be quantified.

Test The 200 test images come from the same edition, except images 187–200, which include colored pages, handwriting, and multi-line cases. The test set shifts thematically toward temple descriptions, historical biographies, and late-imperial political philosophy (out-of-domain).

2.1.2. Dataset B

Training Dataset B consists of 5,000 full-page low-quality images from three sources: illustrated reference books (objects, hierarchies) linked to the Confucian Classics, diagrams and trigram/hexagram lists from the *Yijing*, and maps. Page density varies widely, from heavily populated layouts to nearly empty chapter boundaries where seals often appear, with titles and metadata. The provided JSON annotations specify bounding-box coordinates, class labels (*book-edge*, *image*, *seal*, *text*), and partial transcriptions. Annotation consistency was uneven: (1) **region granularity**: multiple images or text columns were sometimes grouped into a single box, sometimes split individually; (2) **text within diagrams**: embedded text in maps or diagrams was annotated inconsistently. Given these inconsistencies, we **re-annotated 2,308 images** following explicit guidelines (Section 3.3), introducing refined categories (*text-horiz*, *text-one*). Observing that vertical columns were sometimes marked as horizontal lines, we systematically annotated titles, isolated characters, and main columns. This implied a follow-up pipeline: (1) count and order lines (using the line-ordering tool from (Bizais-Lillig et al., 2024) or a YOLO classifier), (2) detect text orientation and rotate, (3) automatically transcribe text.

Test The 200 test images are full-page xylographic documents containing diagrams, object illustrations, and trigram/hexagram lists, closely aligned with the training distribution but excluding maps.

2.1.3. Dataset C

Training Dataset C comprises 5,000 low-quality handwritten line images in regular (*kaishu*) and cursive (*xingshu*) scripts, rotated 90° counterclockwise, drawn mainly from poetry, philosophy, and religion. Annotation presented several challenges: (1) **blanks**: title–author separators, systematically compressed in JSON transcriptions; (2) **punctuation and highlighting marks**: visually identical dots were handled inconsistently—sometimes removed, sometimes converted to modern marks, with semantic semi-commas occasionally added. All such marks were exhaustively removed during initial data treatment to align with evaluation criteria; (3) **repetition markers**: ditto signs were replaced by the repeated character; an experimental placeholder strategy (e.g., "Z") was ultimately deferred to maintain consistency with the official evaluation framework; (4) **character variants**: handwriting introduces graphic simplifications beyond print variants, requiring normalization (Bizais-Lillig, 2025) or special Unicode handling (莊德明, 2001). Transcriptions were inconsistent—faithful to original glyphs in some cases, normalized in others—and a mid-preparation variant-mapping JSON proved unreliable due to internal conflicts.

Test The 200 test images form a homogeneous set—dense lines with minimal blanks, few variants, and no repetition markers or punctuation—focused exclusively on late-19th-century harbor trade, constituting a clear domain shift from the training data.

2.2. External datasets

Four datasets were used to ensure training robustness. The **Guangdong-Hong Kong-Macao Greater Bay Area** competition dataset (gua) consists of 2,000 pages from diverse philosophical and religious sources, characterized by applied transformations including color and contrast variation, rotation, and geometric distortion. **ChiKnowPo** (Bizais-Lillig, 2024) comprises 327 images producing 12,198 line crops in traditional characters, covering diverse genres within a stylistically unified xylographic source; ablation results (Section D) show it complements the `UltraGlyph` synthetic data (Section 3.2). **MTHv2** (Weihong Ma, 2020) aggregates 2,200 pages from Buddhist corpora plus challenging historical documents (Ma et al., 2020) with strong topical and calligraphic diversity. The **Traditional Chinese OCR Synthetic dataset** (Lin, 2025) comprises 4,103,595 line images on Hugging Face, featuring traditional characters in vertical and horizontal orientations with varied calligraphic renderings and simulated degradation, though its modern syntax and restricted vocabulary limit its relevance to historical documents.

2.3. Dataset interoperability

To ensure cross-dataset compatibility, we applied a standardized normalization pipeline to all text line annotations (see Figure 2). For datasets with polygon-based annotations (Guangdong, MTHv2, ChiKnowPo), we extracted the minimal bounding rectangle, applied binary masking to isolate text regions, computed perspective transforms to correct rotation and skew, and applied 90° counterclockwise rotation for consistent orientation. For the synthetic dataset containing pre-rotated vertical text, only 90° rotation was applied. Input formats varied across datasets —JSON arrays, comma-separated coordinate files, and PageXML structures—but all converged to uniform axis-aligned rectangular crops with horizontal orientation, eliminating format-specific biases prior to model training. Table 4 in Appendices summarizes the number of extracted lines.

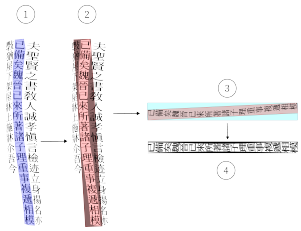


Figure 2: Preprocessing and line-extraction steps to uniformize data format before training

3. Our method

Pipeline Overview: We generate `UltraGlyph` synthetic lines, pretrain OCR models, fine-tune on EvaHan data, and evaluate OCR (Datasets A/C) and layout detection (Dataset B). See Figure 1.

3.1. Data preprocessing

To handle issues in Datasets A and C (see Section 2.1), we developed an orientation classifier $(0^\circ, 90^\circ, 180^\circ, 270^\circ)$ trained on 12,184 images. `ResNet-18` (He et al., 2016) (Adam, $lr = 0.001$, batch 32) reached 99.30% validation accuracy (loss: 0.0298) at epoch 7, outperforming `YOLO11s-cls` (95.0% accuracy). While not integrated into the final production inference, this classifier demonstrates the feasibility of mitigating orientation noise in large-scale historical corpora.

We also prepared a new line classifier (one/multiple line(s)) and a multi-line segmentation detector using Bizais-Lillig et al. (2024), until

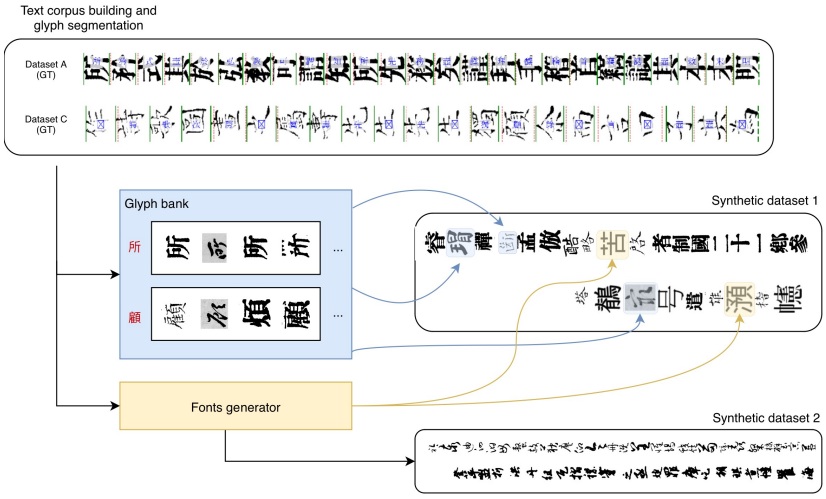


Figure 3: *UltraGlyph* gen. pipeline: glyphs extracted via snap-to-grid segmentation are combined with font-rendered characters, blending authentic crops with synthetic glyphs for max. vocabulary coverage.

the organizers stated that Datasets A and C would contain only single lines and that transcription was not required for Dataset B. Discrepancies identified during visual inspection of the ground-truth annotations necessitated a systematic re-annotation effort to enhance layout detection robustness (see Section 3.3).

3.2. Synthetic dataset for model pre-training (Closed-Task)

We developed `UltraGlyph`, a synthetic data generation pipeline combining digital fonts with authentic glyphs extracted from Datasets A and C. The system prioritizes character-level representation to increase rare character frequency, generating text sequences via stochastic sampling of the ground-truth vocabulary (see Figure 3).

Glyph Extraction and Banking The pipeline builds a library $\mathcal{B}$ by segmenting Datasets A and C. For an image containing n characters, we identify ink boundaries $[x_{start}, x_{end}]$ via vertical projection of the binarized mask M. We implement a *snap-to-grid* strategy: a theoretical grid places boundaries $T_i = x_{start} + i \cdot w$ (with w the mean character width), while physical segments (continuous ink islands) are detected simultaneously. For each character $i \in \{1, \ldots, n-1\}$, the final cut x_i snaps to the nearest physical edge within tolerance:

$$x_i = \begin{cases} p & \text{if } \exists p \in \text{Edges}(M) \text{ s.t. } |p - T_i| < 0.15w \\ T_i & \text{otherwise} \end{cases}$$

where $\text{Edges}(M) = \{x \mid (P(x) > 0 \wedge P(x-1) = 0) \vee (P(x) = 0 \wedge P(x-1) > 0)\}$ and $P(x) = \sum_y M(x, y)$. Regions failing alignment are excluded to avoid fragmented characters.

Hybrid Composition and Stochastic Augmentation Lines are generated by sampling a sequence C from the global vocabulary. To address class imbalance, the sampling distribution is weighted inversely to character frequency, ensuring that rare glyphs are selected with higher probability. For each character $c \in C$, the source is selected with probability p: an image crop from $\mathcal{B}(c)$ ($p = 0.85$) or a TTF-rendered glyph ($p = 0.15$; split is qualitative and should be benchmarked). Each glyph g is resized to height $H \in \{64, 96\}$ with vertical jitter $\delta \sim \mathcal{U}(0, H - h_g)$. The line $L = \text{concat}(g_1, \ldots, g_n)$ is augmented via $I_{synth} = \mathcal{T}(\text{Pad}(L))$, where $\mathcal{T}$ applies morphological filters, Gaussian noise $\mathcal{N}(0, \sigma^2)$, and elastic deformations, generating 134,528 lines.

Font-only Vocabulary Expansion A complementary set of 100,000 clean font-rendered lines prevents bias toward synthesis artifacts. Stroke weight is modulated ($W \in \{\text{bold, normal, thin}\}$ via dilation/erosion kernels $k \in \{2, 3\}$), ink wear simulated (white noise density $\rho \in [0.01, 0.04]$), and character overlap allowed via $C_{x:x+w} = \min(C_{x:x+w}, g)$ to preserve stroke continuity. Final output: $I_{clean} = \text{Otsu}(\text{Blur}(L, \sigma))$.

3.3. Annotation

We re-annotated half of Dataset B to better reflect page scan content, introducing two finer-grained categories —"text-horiz" for horizontal text and "text-one" for single characters—to address the heterogeneity of the broad "text" class (Dataset B⋆ from here on). We also employed oriented bounding boxes to capture objects in non-standard layouts, such as circular character arrangements. For images, overlapping annotations were used when-

ever sub-images were nested within larger ones to preserve structural information, and exhaustive coverage of all categories was ensured (see Figure 4). While these annotations more faithfully represent the data, they increase learning difficulty for layout model (see Table 3).

3.4. Closed modality - Strategy

3.4.1. Tasks A and C

We trained two line-level recognition models. First, a custom `CRNN` (`PyTorch`) combining a convolutional backbone (three GroupNorm conv blocks expanding to 256 filters with max-pooling) and two bidirectional LSTM layers (hidden size 450, dropout), optimized with CTC loss and aggressive `Albumentations` augmentation (blur, noise, contrast, dropout, channel shuffling, embossing, grayscale). Second, `PaddleOCR v5` with `SVTR_HGNet` architecture (PPHGNetV2-B4 backbone, multi-scale SVTR neck), trained with combined CTC+NRTR loss and built-in augmentation (RecAug, RecConAug, RandAugment, multi-scale sampling).

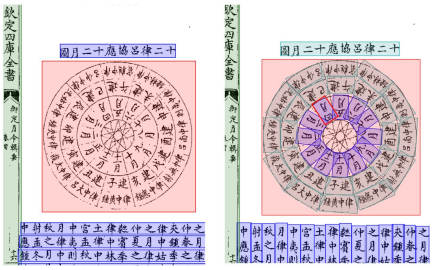


Figure 4: Example of re-annotation in dataset B*.

Both followed a two-stage strategy: (1) foundation-model training on 200,000+ `UltraGlyph` synthetic lines (Section 3.2, using only organizer-provided data); (2) fine-tuning on task-specific datasets, enabling adaptation from synthetic to real documents while retaining broad character coverage.

3.4.2. Task B

For layout detection, we fine-tuned three architectures on the re-annotated Dataset B⋆ (see Section 3.3): **YOLO OBB** (YOLOv8l-obb) for oriented bounding boxes, and **YOLO12s** and **RT-DETR** (large) for axis-aligned bounding boxes. The annotated set (2,308 images, 6 categories; see Table 1) was split 80/20 into train and validation sets (see Table 5 for hyperparameters).

YOLO OBB was trained for 200 epochs on the pre-trained `yolov8l-obb` weights. For YOLO12s, training employed a two-phase schedule (400 epochs total): box loss weight was increased in the first 50 epochs to prioritize localization, then classification weight was raised for the remaining epochs. RT-DETR large was also evaluated as an additional baseline. Full hyperparameter settings are reported in Appendix C.

Label	Count	Label	Count
ext	9,886	text_horiz	3,168
image	3,348	text_one	4,852
book edge	1,825	seal	230

Table 1: Label distribution after re-annotation of Dataset B*.

3.5. Open task - Strategy (tasks A and C)

We evaluated several VLMs including `Qwen 2.5-VL 72B`, `Qwen3-VL 8B and 30B-A3B`,`Gemini 3 Flash`, `Claude Sonnet 4.5` and `GPT-4o`[1]. All models were used with temperature 0 and fixed seed. Each model is governed by a specialized prompt structure (See AppendixA, Figure 7) enforcing a visual-first transcription protocol that explicitly prohibits semantic auto-correction or modern normalization of character forms.

Two dimensions of the in-context learning framework were investigated: the number of shots required for adequate performance, and the effect of exemplar selection strategy. We compared a **random baseline**, in which k shots ($k \in \{5, 10, 25\}$) are drawn stochastically from the training pool, against a **variant-aware dynamic selection** that ranks candidates by their overlap with the non-standard character variants present in the target input, formally defined as $\text{Score}(E) = |V_{\text{target}} \cap V_{\text{candidate}}|$. Together, these two axes allow us to assess each model's sensitivity to both the quantity and contextual relevance of in-context exemplars.

3.6. Metrics

For Tasks A and C, we evaluate OCR quality using CER, NED, and character-level F1. For Task B, we use mAP@[0.5:0.95]. EvaHan defines a single comprehensive score for Tasks A and C:

$$S = 0.5 \cdot (1 - \text{CER}) + 0.3 \cdot \text{F}_1 + 0.2 \cdot (1 - \text{NED})$$

where CER penalizes character-level errors, F_1 balances precision and recall at the character level, and NED penalizes length discrepancies between predicted and reference transcriptions.

[1] Llama 4 Maverick, Mistral Large, Bytedance Seed 1.6 Flash, and InternVL2.5-78B were initially included in the evaluation but produced incomplete results across sampling conditions and were therefore excluded from the final benchmark.

4. Results

4.1. Tasks A and C

Table 2 reports the best-performing configuration per system on Datasets A and C; full per-configuration results and ablation studies are provided in Appendix D and Figure 11.

Closed-modality models perform strongly on Dataset A but degrade markedly on Dataset C. `CRNN-FT` reaches $S_A = 0.963$ (CER $= 0.038$) and $S_C = 0.782$ (CER $= 0.221$), marginally outperforming its foundation model but showing weak transfer to handwriting. `PaddleOCR`, with its multi-scale SVTR neck and CTC+NRTR loss, is more resilient: it trails on Dataset A ($S_A = 0.947$, CER $= 0.056$) but achieves the best closed-modality score on Dataset C ($S_C = 0.852$, CER $= 0.151$), a 31.7% relative CER reduction over `CRNN-FT`, suggesting that multi-head training and multi-scale features confer stronger generalization to cursive strokes.

Model	Norm.	Dataset A				Dataset C			
		Config	Score↑	CER↓	F1↑	Config	Score↑	CER↓	F1↑
Closed modality									
CRNN-FT	norm.	–	**0.9632**	0.0382	0.9666	–	0.7823	0.2212	0.7903
CRNN-FT	raw	–	**0.9350**	0.0663	0.9381	–	0.7513	0.2520	0.7590
CRNN (found.)	norm.	–	0.9623	0.0391	0.9655	–	0.7828	0.2212	0.7916
CRNN (found.)	raw	–	**0.9350**	0.0663	0.9381	–	0.7432	0.2607	0.7517
PaddleOCR	norm.	–	0.9465	0.0558	0.9515	–	**0.8520**	0.1509	0.8565
PaddleOCR	raw	–	0.9191	0.0832	0.9241	–	**0.8130**	0.1900	0.8176
Open modality									
PaddleOCR	norm.	v5	**0.9688**	0.0317	0.9698	v4	0.8927	0.1081	0.8946
PaddleOCR	raw	v5	0.9433	0.0572	0.9443	v4	0.8430	0.1578	0.8448
Gemini 3-Flash	norm.	0s	0.9612	0.0388	0.9612	0s	0.9360	0.0652	0.9384
Gemini 3-Flash	raw	dyn-5	0.9537	0.0466	0.9543	0s	0.9090	0.0923	0.9115
Qwen2.5-VL-72B	norm.	dyn-25	0.9581	0.0430	0.9604	0s	**0.9499**	0.0503	0.9503
Qwen2.5-VL-72B	raw	dyn-25	0.9501	0.0510	0.9523	0s	**0.9276**	0.0727	0.9280
Qwen3-VL-30B	norm.	dyn-25	0.9651	0.0358	0.9669	dyn-10	0.9147	0.0861	0.9162
Qwen3-VL-30B	raw	dyn-25	**0.9632**	0.0376	0.9651	0s	0.8884	0.1125	0.8904
Qwen3-VL-8B	norm.	dyn-5	0.9681	0.0321	0.9684	0s	0.9408	0.0595	0.9415
Qwen3-VL-8B	raw	dyn-5	0.9626	0.0376	0.9629	rnd-5	0.9210	0.0797	0.9226

Table 2: Best OCR performance per model and normalization on Datasets A and C. Each row reports the best-scoring configuration for that model. Closed-modality models have no ICL (–). For open-modality LLMs, **Config** gives the best ICL strategy and shot count across $k \in \{5, 10, 25\}$: *0s* = zero-shot, *dyn-k* = variant-aware dynamic, *rnd-k* = random. For PaddleOCR, the best checkpoint version (v1–v5) may differ between datasets.

In the open modality, `PaddleOCR` v5 (norm.) leads on Dataset A ($S = 0.969$, CER $= 0.032$), confirming that external data yields consistent but modest gains on printed text. On Dataset C, v4 reaches $S = 0.893$, outperforming all closed systems, likely due to broader cursive style coverage.

Among VLMs, `Gemini 3-Flash` and the `Qwen3-VL` family (8B and 30B-A3B) are competitive on Dataset A ($S \approx 0.96$–0.97), on par with dedicated OCR systems. `Qwen3-VL-8B` matches or slightly outperforms its 30B-A3B counterpart on both datasets, suggesting that architecture and Chinese pre-training coverage matter more than parameter count. `Qwen2.5-VL-72B` trails Qwen3 on Dataset A but leads all LLMs on Dataset C at zero-shot ($S = 0.950$), indicating stronger generalization to cursive handwriting. `GPT-4o` and `Claude Sonnet 4.5` lag considerably on both datasets ($S_A \approx 0.81$ and 0.76; $S_C \approx 0.66$ and 0.47), reflecting limited historical Chinese script coverage in their pre-training (see Appendix, Figure 11).

Impact of ICL strategy and shot count. ICL contributions are modest and strongly model-dependent (Figures 5–6). Few-shot prompting produces substantial CER reductions for weaker models — `Claude 4.5` gains ΔCER ≈ 0.09–0.10 on Dataset A and 0.17–0.19 on Dataset C, `Qwen2.5-VL-72B` up to 0.084 on Dataset A — but yields neutral or slightly negative effects for already-strong models (`Gemini 3-Flash`, `Qwen3-VL`), which appear saturated by exemplar injection. Exemplar selection strategy induces only marginal differences ($|\Delta\text{CER}| < 0.015$), with no consistent winner across model families. Taken together, pretraining quality largely supersedes ICL for strong models, while weaker models benefit primarily from script exposure rather than selection strategy.

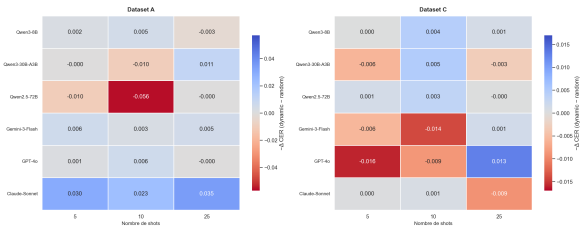


Figure 5: CER differential between Variant-Aware Dynamic and random ICL selection ($\Delta S = S_{\text{dynamic}} - S_{\text{random}}$) across models, datasets, and shot counts. Blue: dynamic outperforms random; red: the reverse.

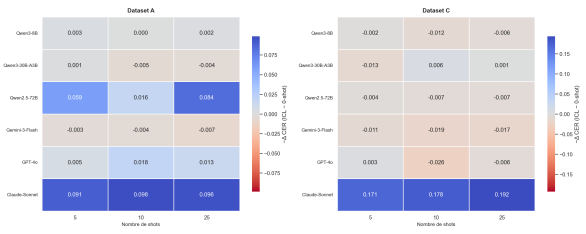


Figure 6: CER differential of ICL over zero-shot inference ($\Delta S = S_{\text{ICL}} - S_{\text{0shot}}$) per model, dataset, and shot count. Blue cells indicate that zero-shot outperforms few-shot ICL.

Taken together, these results indicate that for strong models the quality of pre-training largely su-

persedes the contribution of in-context exemplars, while weaker models benefit primarily from exposure to the target script rather than from the selection strategy itself.

Qualitative Assessment See Table 8 and Figure 9 in appendix for a qualitative display. Open models are considerably more resilient to variation and damage. In Dataset A, images deviating from standard printed editions may concentrate errors (a_040, a_190, possibly due to manuscript-style handwriting) or hinder recognition altogether (a_194, possibly due to color interference). A similar pattern appears in Dataset C (c_189), with additional phenomena: punctuation inserted by the closed model, likely from training data contamination, and heightened sensitivity to poorly formed characters such as 月. When horizontal page strokes could plausibly belong to a character, the closed model tends to incorporate them, producing spurious readings (c_175). Both models, as well as human annotators, are further challenged by character variants—including simplified forms—within frequent classes such as 關, 税, 内, 两, and 并. When errors occur, the model typically identifies sub-graphemic components of the target character, as illustrated by the 該/設 confusion in c_150. Both models generally succeed in isolating individual characters, even in Dataset C where density varies and peripheral strokes risk corrupting segmentation. While the open model clearly outperforms the closed model on Dataset C, some of its errors may stem from large-scale modern Chinese training data, as suggested by a possible translation error in c_003 and a character insertion in c_172. The open model also shows reduced robustness when recognition is repeated.

4.2. Task B

Table 3 reports per-class detection performance. **YOLO12s** outperforms **RT-DETR** on all aggregate metrics (mAP50: 0.750 vs. 0.723; mAP50-95: 0.486 vs. 0.463). Both models perform strongly on well-represented, visually distinct categories (*book_edge*, *seal*: mAP50 ≥ 0.99), but degrade on *text-horiz* (mAP50: 0.402 and 0.339). `YOLO12s` trained on original annotations scores higher (mAP50: 0.861, mAP50-95: 0.578), but this is partly misleading: those annotations are less exhaustive for textual elements (Fig. 10), so the apparent gain may reflect lower annotation expectations rather than genuine improvement.

`YOLO OBB`, fine-tuned on the reannotated dataset, outperforms both models on aggregate metrics with notable gains on textual classes, suggesting that oriented bounding boxes better capture textual layout patterns and that exhaustive reannotation benefits detection.

Class	P	R	mAP50	mAP50-95
YOLO12s				
All	0.791	0.731	**0.750**	**0.486**
book_edge	0.988	0.991	0.994	0.760
image	0.890	0.714	0.780	0.477
seal	0.939	1.000	0.995	0.782
text	0.881	0.611	0.787	0.444
text-horiz	0.443	0.488	0.402	0.212
text-one	0.606	0.581	0.543	0.240
all “text”	0.764	0.591	0.684	0.366
YOLO OBB				
All	0.785	0.809	0.808	0.583
book_edge	0.491	0.664	0.443	0.050
image	0.893	0.880	0.894	0.619
seal	0.911	0.976	0.982	0.943
text	0.822	0.777	0.864	0.652
text-horiz	0.759	0.753	0.803	0.631
text-one	0.834	0.801	0.859	0.603
RT-DETR large				
All	0.794	0.707	0.723	0.463
book_edge	0.969	0.979	0.992	0.723
image	0.867	0.647	0.740	0.475
seal	0.964	1.000	0.995	0.748
text	0.875	0.609	0.751	0.431
text-horiz	0.440	0.446	0.339	0.180
text-one	0.648	0.560	0.521	0.224
YOLO12s on original dataset				
All	0.877	0.810	0.861	0.578
book_edge	0.921	0.970	0.983	0.685
image	0.776	0.667	0.720	0.474
seal	0.953	1.000	0.995	0.759
text	0.858	0.602	0.747	0.396

Table 3: Per-class detection results on Dataset B and Dataset B* validation set.

5. Conclusion

We presented a complete pipeline for EvaHAN 2026, combining `UltraGlyph` synthetic data (234,528 lines: 134,528 hybrid, 100,000 font-only), two-stage OCR pretraining, and re-annotation of 2,308 Dataset B images with 6 refined categories and oriented bounding boxes. Closed-modality CRNN-FT reaches 96.3% on Dataset A but degrades to 78.2% on Dataset C; PaddleOCR proves more resilient (94.7% / 85.2%). In open modality, PaddleOCR v5 leads on printed text (96.9%) while `Qwen2.5-VL-72B` zero-shot leads on cursive (95.0%). ICL yields marginal gains for strong models (ΔCER < 0.015) but substantial improvements for weaker ones (`Claude 4.5`: +17–19% on Dataset C); variant-aware selection shows no consistent advantage over random sampling. These results confirm that pretraining quality and data coverage dominate performance, with ICL primarily benefiting underspecified models.

6. Acknowledgements

We thank the Evahan2026 organizing team for providing the shared task and the associated data. This study was conducted as part of the DH master's program at École nationale des chartes–PSL[2]. It also received support from the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001). We used Claude AI to help condense the paper and proofread the English. Any remaining mistakes are our own.

7. Code and models availability

All codes and models are available on https://github.com/Bizais-Lillig/enchanteam.

8. Bibliographical References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Marie Bizais-Lillig. 2025. Transcrire des éditions impériales chinoises pour acquérir un corpus numérique : Questions, choix et réflexions. In Alix Chagué, Thibault Clérice, and Ariane Pinche, editors, *Apprendre à Lire aux Machines*, pages 121–128.

Marie Bizais-Lillig, Chahan Vidal-Gorène, and Boris Dupin. 2024. Optimizing HTR and Reading Order Strategies for Chinese Imperial Editions with Few-Shot Learning. In *Document Analysis and Recognition – ICDAR 2024 Workshops*, volume 14936 of *Lecture Notes in Computer Science*, pages 37–56. Springer Nature Switzerland.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901.

Bo Chang, Qiong Zhang, Shenyi Pan, and Lili Meng. 2018. Generating handwritten chinese characters using cyclegan. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 199–207. IEEE.

Yan Hon Michael Chung and Donghyeok Choi. 2025. Finetuning vision-language models as ocr systems for low-resource languages: A case study of manchu.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. 2025. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*.

S. Gupta, T. Xie, and D. Roth. 2023. Coverage-based example selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 13924–13935.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.

G. Kim, T. Hong, M. Yim, J. Nam, J. Park, and S. Kim. 2022. OCR-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, pages 498–517.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*.

Y. Li, Y. Li, Y. Wang, and L. Wang. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR 2022)*.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

[2] https://www.chartes.psl.eu/formations/masters/master-humanites-numeriques

10435–10492, Vienna, Austria. Association for Computational Linguistics.

Y. Luo, Z. Chen, Y. Wang, and Z. Zhang. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Transactions of the Association for Computational Linguistics*, 12:1–24.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization.

Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 2025. Churro: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34765–34812.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. 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*, 39(11):2298–2304.

Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. 2021. A modular and automated annotation platform for handwritings: evaluation on under-resourced languages. In *International Conference on Document Analysis and Recognition*, pages 507–522. Springer.

Chahan Vidal-Gorène, Bastien Kindt, and Florian Cafiero. 2026. Under-resourced studies of under-resourced languages: lemmatization and pos-tagging with llm annotators for historical armenian, georgian, greek and syriac. *arXiv preprint arXiv:2602.15753*.

Dongbo Wang and Dongmei Zhu. 2025. Benchmarking the ancient books capability of multimodal large language models. *npj Heritage Science*, 13(1):339.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2020)*, pages 1192–1200.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, et al. 2025. Benchmarking vision-language models on chinese ancient documents: From ocr to knowledge reasoning. *arXiv preprint arXiv:2509.09731*.

Yuyi Zhang, Yongxin Shi, Peirong Zhang, Yixin Zhao, Zhenhua Yang, and Lianwen Jin. 2025. Megahan97k: A large-scale dataset for mega-category chinese character recognition with over 97k categories. *Pattern Recognition*, 167:111757.

Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. 2025. Xunzi-mllm: a multimodal large language model for ancient text and image recognition. *Digital Scholarship in the Humanities*, 40(2):709–722.

Zhuang Deming 莊德明. 2001. Zhongwen diannao quezi jiejue fang'an 中文電腦缺字解決方案 (report on a method to solve the lack of characters on computers in chinese language). In 第一屆中國文字學國際學術研討會 *(The First International Conference on the Study of Chinese characters*.

9. Language Resource References

Bizais-Lillig, M. 2024. *HTR ground-truth of Chinese xylographic editions*. Document Analysis and Recognition –ICDAR 2024 Workshops. ICDAR 2024. Lecture Notes in Computer Science (v1.0.0, p. 37-56), 1.0. PID /10.5281/zenodo.14452717.

Lin,Zih-Ci. 2025. *Traditional Chinese OCR Synthetic*.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, Yongpan Wang. 2020. *Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization*.

Appendices

A. Prompt for ICL

System Prompt (Paleographic Persona):
You are a professional paleographer specializing in Chinese. Your goal is a character-for-character high-fidelity transcription.
Rules:
1. DO NOT CORRECT.
2. NO SEMANTIC ADAPTATION.
3. VISUAL PRIORITY.
4. Output ONLY characters.

Few-Shot Examples (k iterations):
User: Transcribe this image: [Image_i]
Assistant: [Gold_Transcription_i]

Target Input:
User: Transcribe this text. Output only the transcription. [Target_Image]

Figure 7: Schematic representation of the interleaved multi-modal prompt structure.

B. Lines used per dataset

Dataset	Total	Ratio	Used
Dataset A	5,000	100%	5,000
Dataset C	5,000	100%	5,000
ChiKnowPo	28,551	100%	28,551
Guangdong	50,014	45%	22,506
MTHv2	105,579	85%	89,742
Traditional Synth.	50,000	60%	30,000
UltraGlyph (clean)	100,000	85%	85,000
UltraGlyph (synth.)	134,528	50%	67,264
Total	**478,672**	—	**333,063**

Table 4: Line extraction and sampling configuration across all datasets.

C. YOLO Hyperparameters

Parameter	Epochs 1–50	Epochs 51–400
batch / imgsz	8 / 1024	8 / 1024
patience	30	30
dropout / iou	0.3 / 0.6	0.3 / 0.6
cos_lr / half	True / True	True / True
multi_scale	0.15	0.15
box loss weight	8.5	7.5
cls loss weight	1.5	2.5
hsv_s / hsv_v	0.8 / 0.8	0.8 / 0.8
degrees / fliplr / scale	0.2 / 0.3 / 0.1	0.2 / 0.3 / 0.1

Table 5: YOLO training hyperparameters.

D. Ablation Study: Incremental Training Data for PaddleOCR

We conducted a cumulative ablation adding training corpora incrementally across five configurations (Table 6, Figure 8).

Config	Added corpus	Dataset A		Dataset C	
		Score	CER	Score	CER
v1	Official only + UltraGlyph	0.772	0.232	0.681	0.324
v2	+ ChiKnowPo	0.963	0.038	0.870	0.130
v3	+ Guangdong	0.963	0.038	0.860	0.141
v4	+ MTH	0.969	0.031	**0.893**	**0.108**
v5	+ SynthHF	**0.969**	**0.032**	0.871	0.130

Table 6: Ablation results for open PaddleOCR (normalized). Each configuration accumulates the previous corpora.

The dominant gain is produced by the transition from v1 to v2: adding ChiKnowPo reduces CER from 0.232 to 0.038 on Dataset A and from 0.324 to 0.130 on Dataset C, a drop of over 85% and 60% respectively.

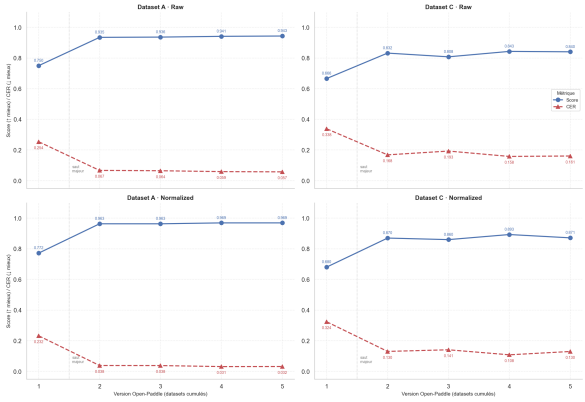


Figure 8: Incremental performance of open PaddleOCR configurations on Datasets A and C (normalized). Each configuration adds one external corpus to the previous, following the order described in the text.

Subsequent additions yield more nuanced effects: Guangdong (v3) leaves Dataset A unchanged while marginally degrading Dataset C (S : $0.870 \rightarrow 0.860$); MTH (v4) peaks on both datasets ($S_C = 0.893$, $S_A = 0.969$); SynthHF (v5) maintains Dataset A but reverts Dataset C to near-v2 levels ($S = 0.871$). Beyond v2, gains are modest and non-monotonic, with optimal configurations differing: v4 for Dataset C, v4–v5 (tied) for Dataset A.

E. Layout predictions

Figure 10 displays layout predictions with and without re-annotation.

F. Performance Barplots

Figure 11 presents the full per-configuration score distributions for all models, datasets, and modalities, ordered by comprehensive score (normalized evaluation condition).

G. Character Recognition Error Types in Dataset A and Dataset C

Across both open- and closed-modality models, character recognition errors (CR) dominate: 63.8% of all errors in open-modality (1,689/2,647) and 66.3% in closed-modality (126/190), as illustrated in Figure 9. This consistent pattern across architectures confirms that character-level recognition is the primary bottleneck, rather than insertion, deletion, or truncation.

G.1. Character Variant Effects in Dataset A and Dataset C and Potential OCR Improvements

Metric	Value
Total characters analyzed	40,388
Correct characters	33,846
Variant-character matches	2,496
True recognition errors	4,046
Variant share of non-exact differences	38.15%

Table 7: Variant-based breakdown of character recognition outcomes in the evaluated datasets. A substantial portion of non-exact differences arises from character variants rather than true OCR recognition errors.

Incorporating variant normalization could potentially improve the combined recognition accuracy of open- and closed-modality models on Dataset A and Dataset C by up to 38.15% beyond the current baseline (Table 7). To assess this, we used the variant.json file provided by the EVAHAN benchmark organizers to examine whether character recognition errors were caused by orthographic variation. Ambiguous many-to-many mappings were removed, retaining only cases where one standard character corresponds to one or multiple variant forms. Using this cleaned mapping, we rechecked 1,813 character recognition errors from both modalities. In total, 40,388 characters were examined: 33,846 exact matches, 2,496 variant matches, and 4,046 genuine recognition errors. Variant matches account for 38.15% of all non-exact differences.

Table 8 displays qualitative error analysis for selected examples from the test datasets.

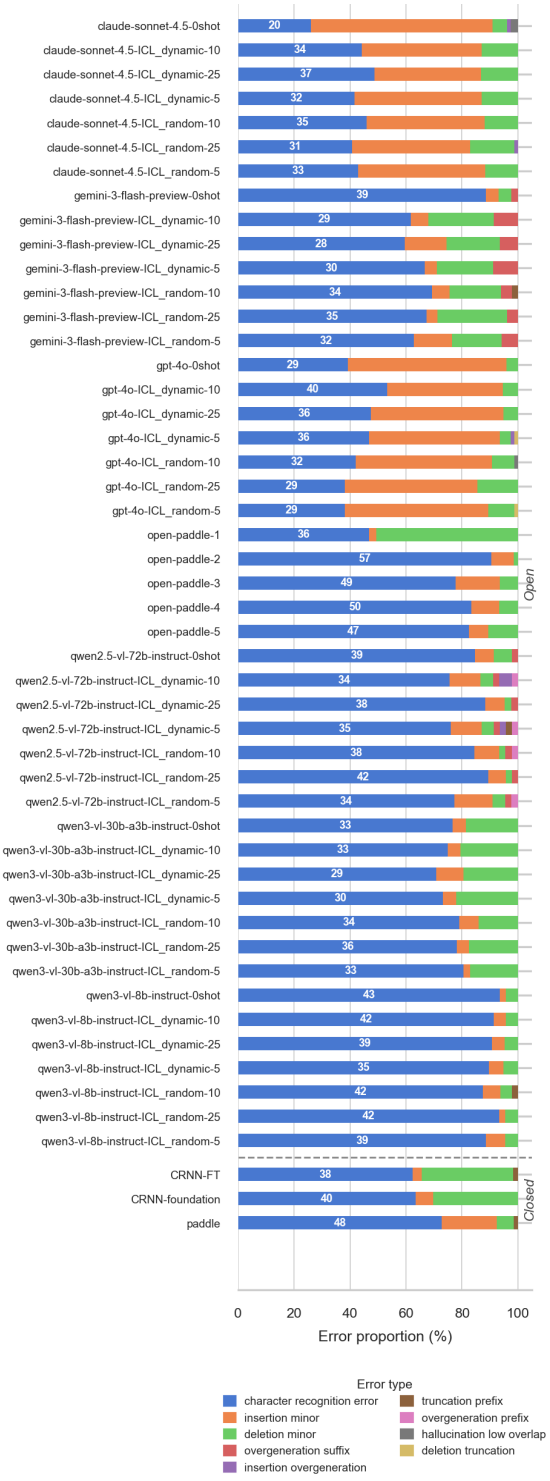


Figure 9: OCR Error Type Distribution Across Models and Datasets.

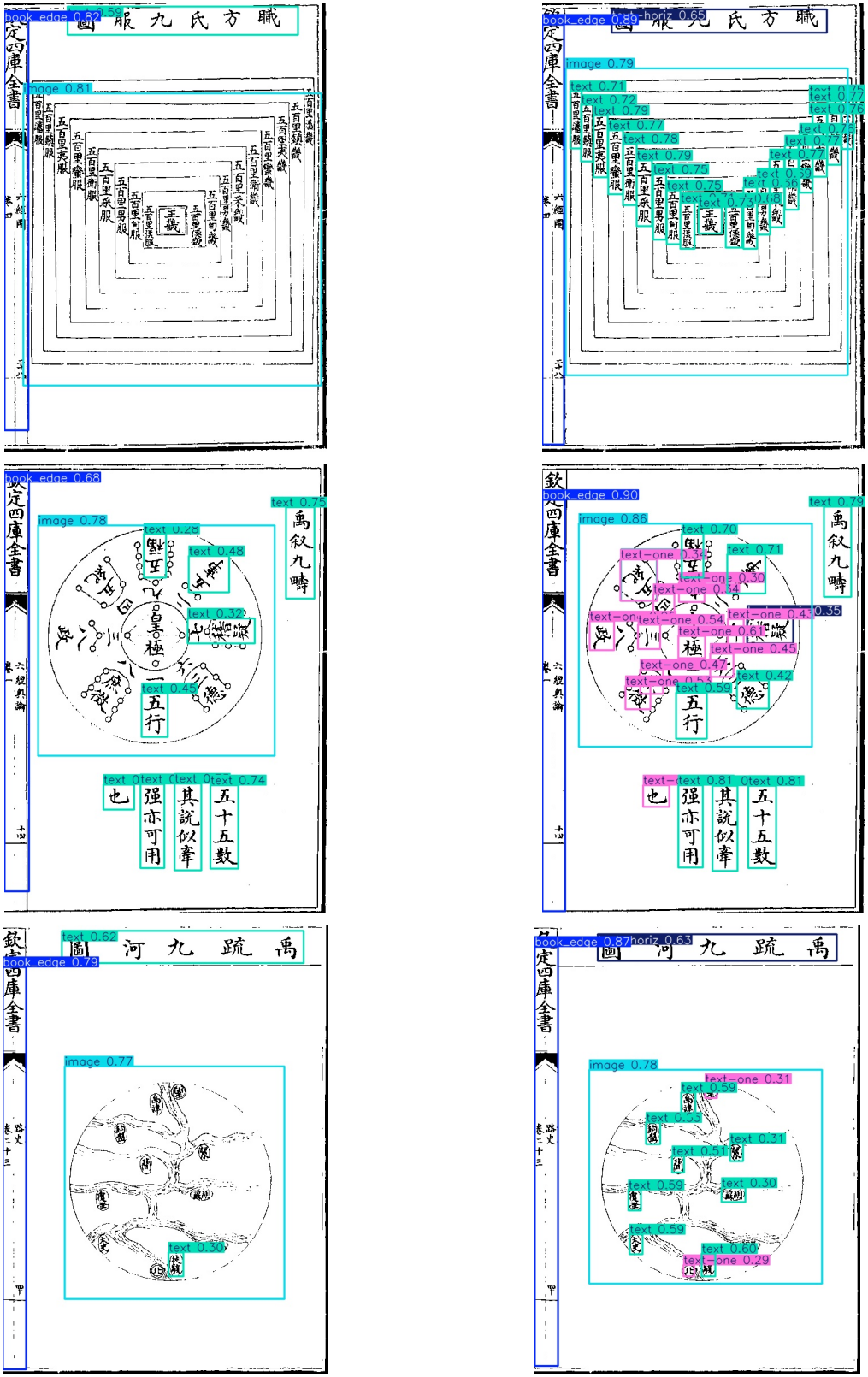


Figure 10: Predictions using YOLO12s model. On the left, trained on original annotations. On the right, train using the Dataset B⋆ annotations.

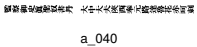
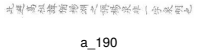
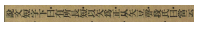
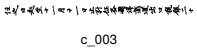
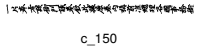
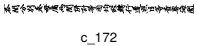
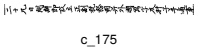
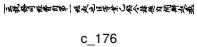
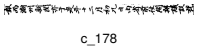
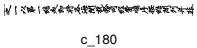
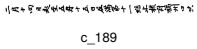
Image	Model	Text
		Dataset A
 a_040	**GT**	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	CRNN-FT (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	open-paddle-5 (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	qwen3-vl-30b (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
 a_190	**GT**	此避高祖諱猶楊淵之稱楊泉非一字泉明也
	CRNN-FT (CER: 22.2%)	此避高祖諱齋楊調之楊泉非一字矣明也
	open-paddle-5 (CER: 0.0%)	此避高祖諱猶楊淵之稱楊泉非一字泉明也
	qwen3-vl-30b (CER: 5.6%)	此避高祖諱猶楊淵之稱楊泉非一字泉明也
 a_194	**GT**	說文短字下曰有所長短以矢爲正從矢豆聲段氏曰當云
	CRNN-FT (CER: 65.2%)	說文短下有所曇矢爲
	open-paddle-5 (CER: 17.4%)	說文短字下曰有所長短以矢為正从矢豆聲 段氏曰 當云
	qwen3-vl-30b (CER: 8.7%)	說文短字下曰有所長短以矢為正从矢豆聲段氏曰當云
		Dataset C
 c_003	**GT**	任之日起至十一月十一日止計征各國洋商進出口稅銀二十
	paddle (CER: 12.0%)	任之日赳至十一月十一日止計征公國洋商道出口稅銀二十
	qwen2.5-vl-72b (CER: 8.0%)	任之日起到十一月十一日止計征各國洋商進出口稅銀二十
 c_150	**GT**	一片奉旨該衙門議奏欽此據原奏內稱前准總理各國事務衙
	paddle (CER: 12.0%)	一片奉者設衙門議奏欽此據原奏內稱前准總理谷國事務衙
	qwen2.5-vl-72b (CER: 8.0%)	一片奉有該衙門議奏欽此據原奏內稱前准總理各國事務衙
 c_172	**GT**	各関分別奏毋庸兩関併計等因均經轉行遵照臣等查粤海関
	paddle (CER: 20.0%)	谷関分別奏母庸兩関并計等因均經轉行遵照臣等查粵海閣
	qwen2.5-vl-72b (CER: 12.0%)	各関分別奏毋庸兩関併計等因均經轉行遵照臣等查粵江海關
 c_175	**GT**	二十九日開辦即設立正副稅務司并外國寫字及杆子手通事
	paddle (CER: 28.0%)	三十兀日開辨即設立正副稅務司，并外國冩字及許子手通畫
	qwen2.5-vl-72b (CER: 4.0%)	二十九日開辦即設立正副稅務司并外國寫字及杆子手通事
 c_176	**GT**	云稅務司經費自第一結起也臣等業已飭令赫德自開辦起截
	paddle (CER: 36.0%)	云。稅務司經帶自第一始走也佳等業已。飭令赫德自開辨超，截
	qwen2.5-vl-72b (CER: 8.0%)	云稅務司經費自第一結起也目等業已飭令赫德自開辦起截
 c_178	**GT**	報而潮州新関亦于是年十二月初九日由省前往開辦議設稅
	paddle (CER: 12.0%)	報而潮州新関亦于是平十二月初九日由省前往開辨議設祝
	qwen2.5-vl-72b (CER: 8.0%)	報而潮州新閩亦于是年十二月初九日由省前往開辦議設稅
 c_180	**GT**	之一以第一結起即將各海関稅務司經費順序按結開列并非
	paddle (CER: 12.0%)	之一以第一結起即將谷海関祝務司經當順序按結開列并非
	qwen2.5-vl-72b (CER: 8.0%)	之一以第一結起即將各海関稅務司經費順序按結開列并非
 c_189	**GT**	二月十四日起至五月十五日屆滿第十一結止核計福州口共
	paddle (CER: 20.0%)	二月十四日起至五舟十五日履滿第十一結止，核計稿州口其
	qwen2.5-vl-72b (CER: 4.0%)	二月十四日起至五月十五日在滿第十一結止核計福州口共

Table 8: Qualitative error analysis on Datasets A and C. Blue: character variants; yellow: variant-standard mismatches; red: recognition errors.

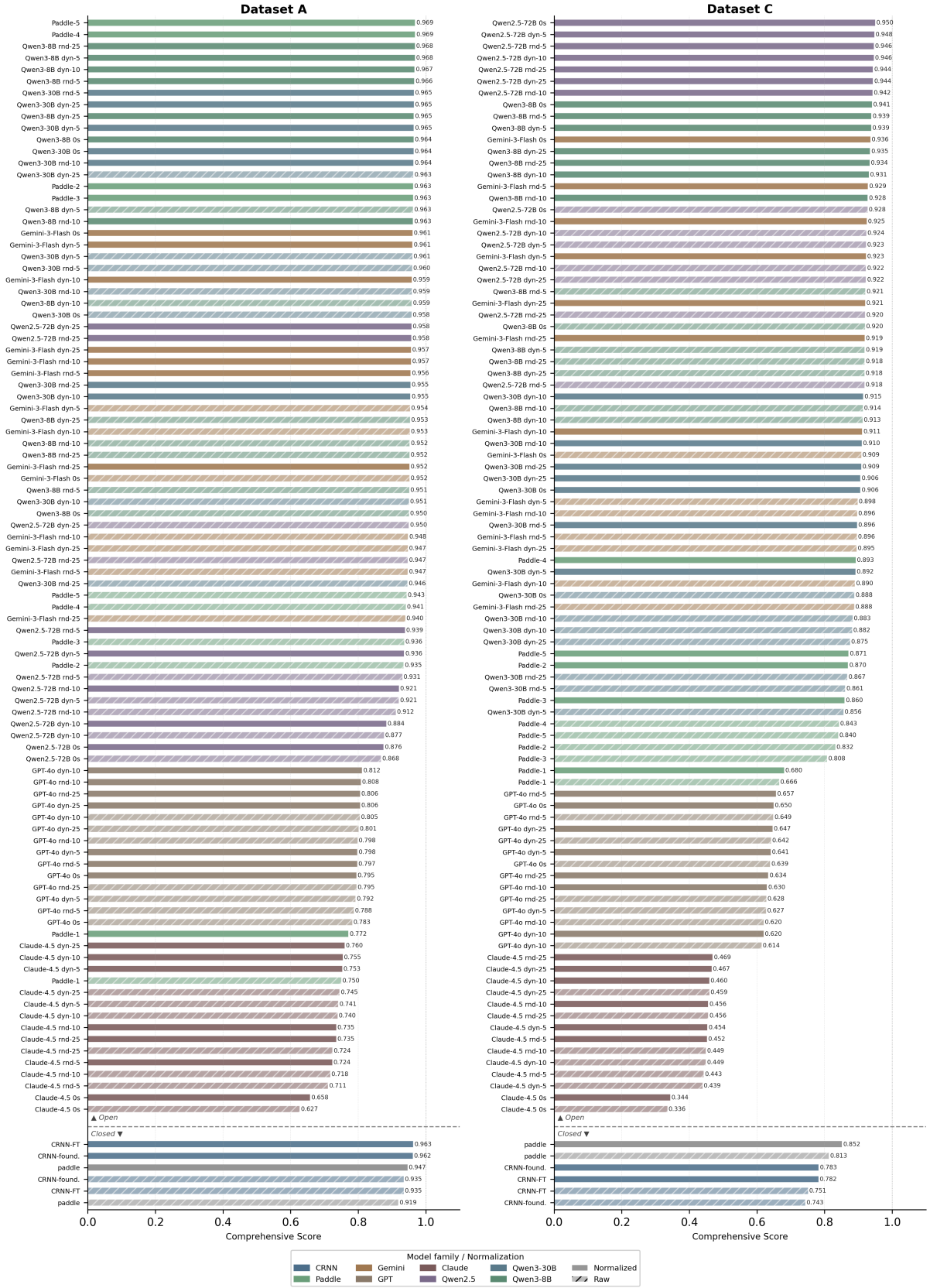


Figure 11: All results, ordered by score