MCHDoc: A Comprehensive Benchmark for Reading Multi-Carrier Chinese Historical Documents

Supplementary Material

In this Supplementary Material, we provide:

- The Details of MCHDoc
- Case Study
- An Exploratory Investigation into Knowledge-Based LLM Post-Correction
- Limitation and Future Work

1. The Details of MCHDoc

1.1. Explanations of six Carriers

AncientBook. AncientBook is primarily written or printed on hemp paper, which is made from processed plant fibers such as hemp, mulberry, or ramie. This type of paper is characterized by its strong and durable texture, good ink absorbency, and resistance to decay. However, after centuries of preservation, many pages exhibit yellowing, ink bleeding, and surface degradation, which complicate character recognition and restoration.

JianDu. JianDu, also known as bamboo slips, was made from bamboo, which was used as writing materials before the widespread use of paper. The vertical arrangement of slips and the uneven surface of bamboo make the text segmented and difficult to align, leading to increased OCR difficulty.

Calligraphy. Calligraphy works are often written on xuan paper. Xuan paper, also known as rice paper, is a high-quality handmade paper traditionally produced in Xuancheng, Anhui Province. It is made from a blend of Pteroceltis tatarinowii bark and straw or other plant fibers. The paper is lightweight yet tough, featuring a smooth surface and excellent ink absorbency, which allows brush strokes to spread naturally and produce rich tonal variations.

Inscription. Inscription refers to characters carved on stone or metal surfaces, such as steles, tablets, or bronze vessels. The rigid and reflective materials produce strong lighting variations and shadows in captured images. Over centuries of exposure and rubbing, these carriers exhibit diverse types of degradation, including surface erosion, cracks, discoloration, and loss of carved depth. Such degradation results in blurred or partially missing characters, making recognition and restoration particularly challenging for vision-based models.

Silk. Silk was used for writing and painting before paper became widespread, particularly during the Warring States. The soft and glossy texture of silk provides a smooth writing surface but tends to age poorly, resulting in cracks, discoloration, and partial loss of text. Moreover, the scripts on silk, such as early clerical or proto-regular styles, differ markedly from the standardized characters used after the Han dynasty.

Oracle Bone. Oracle Bone is the earliest known carrier of Chinese writing, carved on tortoise shells or animal bones during the Shang dynasty. The ancient script forms on oracle bones are drastically different from the later clerical and regular scripts standardized after the Han dynasty, with complex pictographic structures and many extinct glyphs.

Table 1. Statistics of Different Carriers in MCHDoc

Carrier	Scale	Avg. Size	Avg. Character	Categories
AncientBook	3,000	2327×2039	577	15249
JianDu	3,000	278×1834	14	1671
Calligraphy	3192	1211×1647	38	4508
Inscription	5152	691×1594	15	4369
Silk	881	114×162	-	535
Oracle Bone	499	77×135	-	100

1.2. Long-tail Distribution of Four Carriers

In this section, we observe that all four page-level carriers exhibit a clear long-tail distribution, as illustrated in Fig. 1. This observation could guide future efforts to improve recognition performance on rare categories.

1.3. MCHDoc Trainset

In this paper, we introduce a dual-granularity dataset specifically designed for training small-parameter multimodal models. We hypothesize that training data incorporating both broad global context (page-level) and fine-grained local features (character-level) can effectively compensate for the performance degradation often associated with limited model capacity. The dataset is composed of two main parts: (1) **Page-level Data:** it contains 10,000 full document images across four carriers. This part is utilized for training the model's page-level recognition. (2) **Character-level Data:** it contains 6,134 character-level images across six carriers. This collection was curated to train the model's character-level recognition.

1.4. Benchmark Version

The experimental results of Calligraphy reported in the main paper are based on an earlier release of CalliBench (released on June 25th).

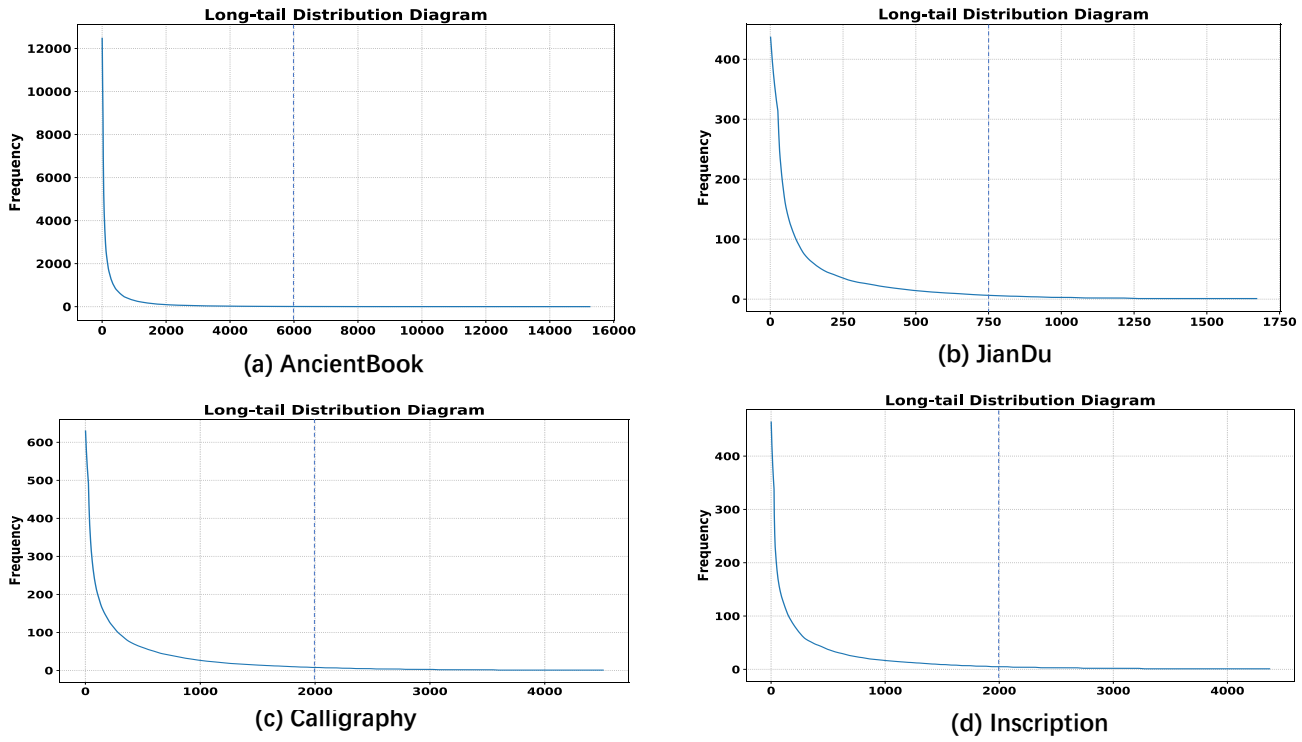

Figure 1. Long-tail Distribution of four Page-level Carriers in MCHDoc (The portion to the right of the blue dashed line corresponds to categories whose occurrence frequency is fewer than five times)

2. Case Study

After extensive evaluations, we observe that the overall performance of existing models still falls short of handling the diverse challenges present in our benchmark. To shed light on these limitations, this section presents a detailed analysis of representative failure cases, which helps reveal the underlying error patterns and provides actionable directions for future research.

For the recognition task, we select representative failure cases from two perspectives. For the page-level setting, we focus on images whose 1-NED falls below 10%. For the character-level setting, we examine samples with an accuracy of 0. All examples are drawn from the best-performing four models in each category, ensuring that the analyzed errors reflect intrinsic challenges rather than deficiencies of weaker systems.

2.1. Effect of Hard Font Style on Recognition Accuracy

Through the analysis of these failure cases, we further summarize several categories of inherently difficult script styles, including Oracle Bone Script, Seal Script, Semi-cursive Script (Xing Shu), and Cursive Script (Cao Shu). These scripts consistently exhibit significantly lower recognition accuracy across models (Fig. 2).

The difficulty largely stems from their high visual complexity, such as irregular glyph structures, large intra-class variability, and strong stylistic deformations. In Oracle Bone Script, the pictographic obscures the correspondence between glyphs and their modern forms. Seal Script introduces dense, rounded, and often symmetric patterns that many models mistake for visually similar characters. Semi-cursive and cursive scripts further challenge the system due to stroke merging, ligatures, unstable writing speed, and personalized handwriting styles, which dramatically reduce the clarity of character boundaries.

Together, these factors make the four script styles particularly error-prone, revealing not only the limitations of current models but also the necessity of designing recognition systems that are more robust to structural variability, historical script diversity, and extreme stroke deformation.

2.2. Degradation Artifacts Impair Recognition Performance

Beyond script complexity, we further identify several forms of document degradations that substantially impair recognition quality (Fig. 3).

Blurred Text often appears in low-resolution scans or due to ink diffusion on aged paper. Such blurring erases fine stroke boundaries and prevents the model from capturing subtle radical-level distinctions, resulting in frequent mis-

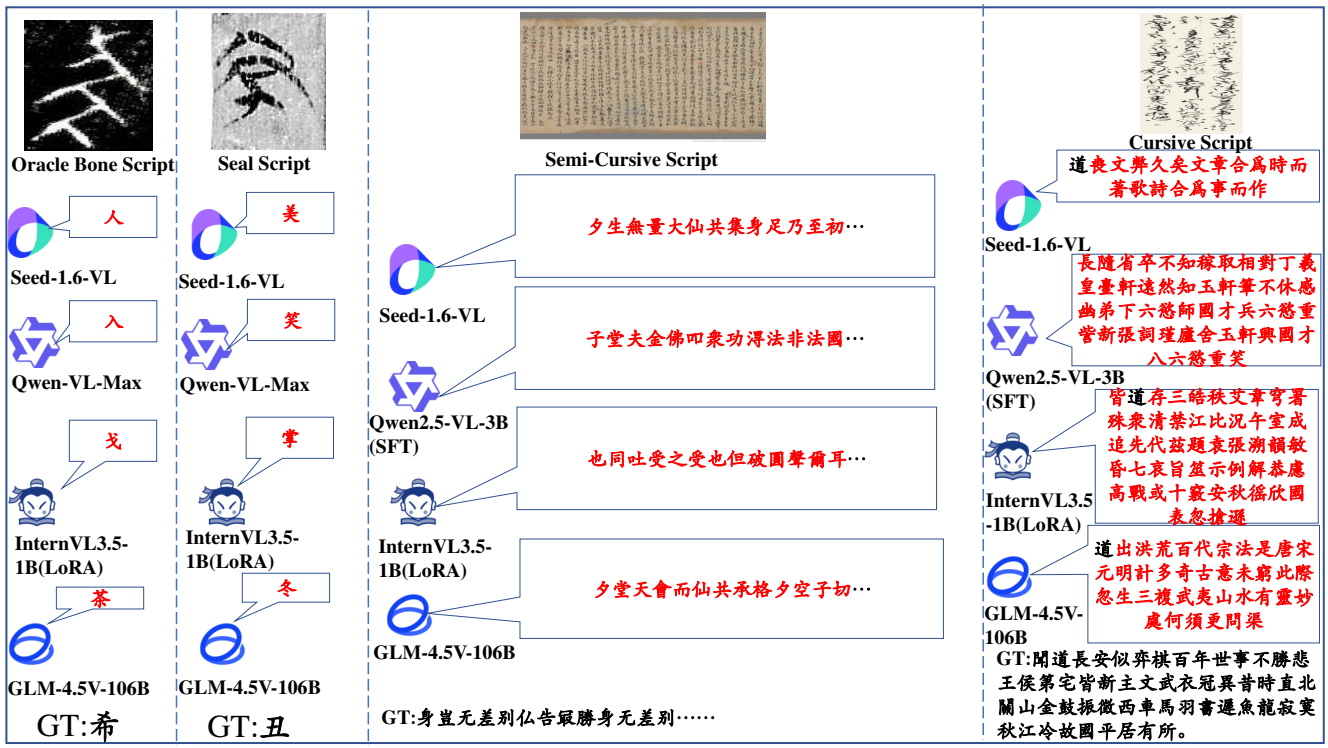

Figure 2. Effect of Hard Font Style on Recognition Accuracy(**Red**: wrong characters)

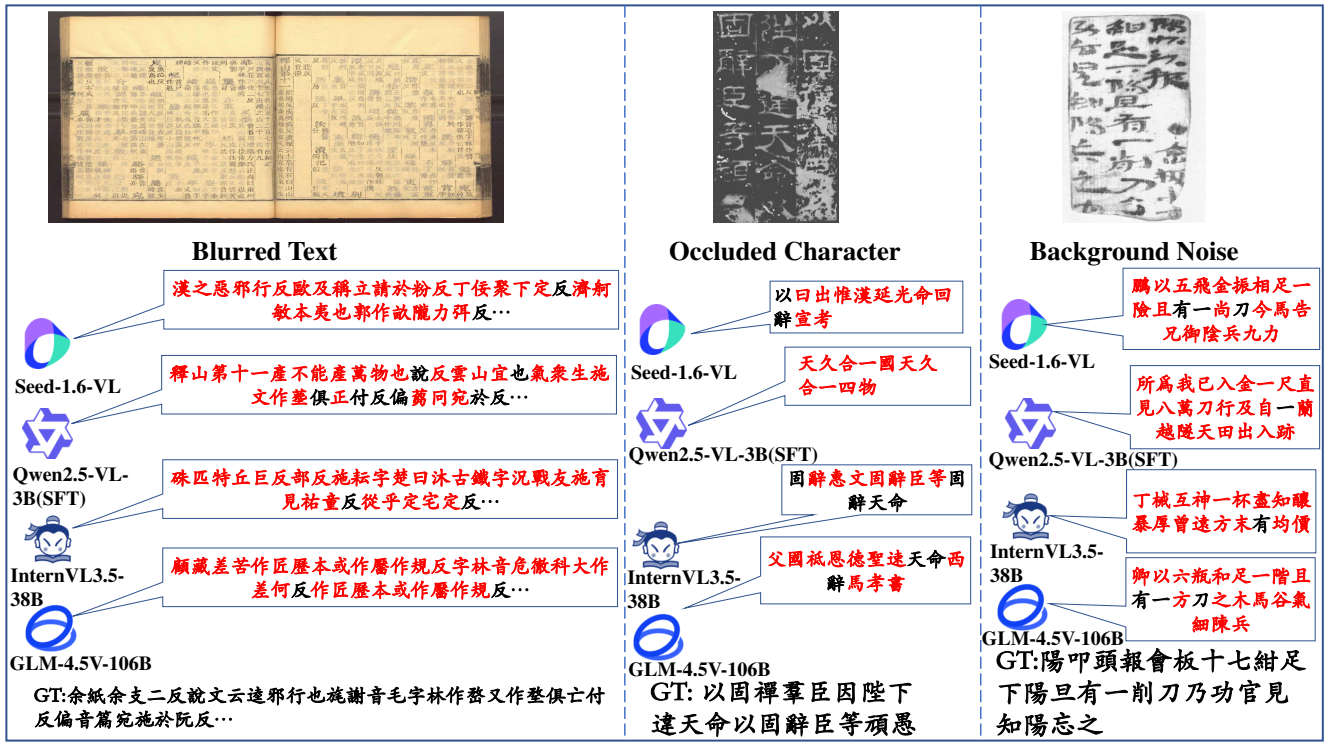

Figure 3. Degradation Artifacts Impair Recognition Performance(**Red**: wrong characters)

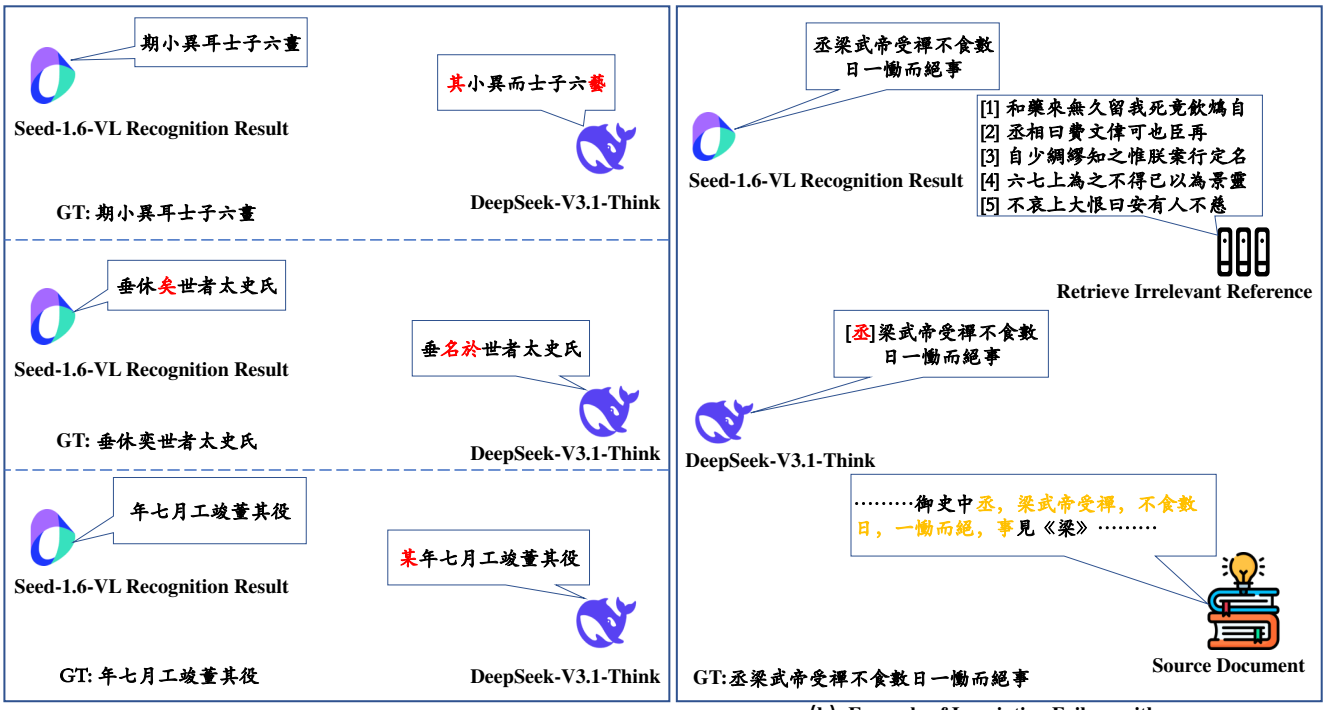

Figure 4. Case of Post-Correction Task (**Red**: wrong characters, **Yellow**: the position of the fragment in the original document)

recognition among visually similar characters.

Occluded Characters, particularly common in Inscription, arise from surface erosion, rubbing artifacts, and mottled damage accumulated over centuries. These occlusions may remove essential stroke segments or distort the overall glyph shape. When critical radicals are partially missing, the model is unable to reconstruct the intended character.

Background Noise mainly refers to irregular white-noise patterns, uneven fiber structures, and scanning artifacts, which overlay the true character strokes. These high-frequency noise components interfere with feature extraction and make the characters appear faint, patchy, and structurally incomplete. As a result, recognition models frequently fail to distinguish genuine strokes from noise-induced patterns, leading to substantial drops in character-level accuracy.

2.3. Over-Correction Phenomena in LLM Post-Correction

In our study, we observed a recurring issue in LLM-based post-correction(Fig. 4(a)): the model occasionally alters text that was originally correct, a phenomenon we refer to as over-correction. This occurs when the model, in its attempt to improve or normalize the output, misinterprets correctly recognized characters or words as erroneous. As a result, the post-corrected output may introduce new errors, reducing overall accuracy despite the intention to improve it.

A primary cause of over-correction is the limited semantic information in OCR outputs. When an LLM attempts post-correction under these conditions, it often lacks sufficient context to reliably distinguish correct from incorrect characters. Consequently, the model relies heavily on its own learned priors and guesses, which can lead to unnecessary or incorrect modifications.

2.4. External Knowledge Leads to Greater Correction Errors in Inscription

Inscriptions are typically partial reproductions of large steles, meaning that each document contains only a small fraction of the text. When using an external knowledge base to assist LLM post-correction, this partiality poses a significant challenge. Queries based on these limited text segments often retrieve references that are irrelevant to the local context of the inscription. Consequently, when the model attempts to leverage these unrelated references for correction, it may introduce additional errors rather than improving accuracy, as shown in Fig. 4(b). This issue illustrates a critical limitation of knowledge-augmented post-correction for inscriptions: the misalignment between the local textual fragment and the global knowledge base can amplify over-correction phenomena. The model's tendency to rely on plausibility cues from irrelevant references can

Table 2. Average Recognition Accuracy (%) on Page-level Benchmark. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	Version	AVG		
		AR	CR	1-NED
Closed-Source MLLMs				
GPT-4o [7]	2025-03-26	13.95	18.31	13.09
GPT-5 [7]	2025-08-07	24.97	26.48	24.84
Doubao-Seed-1.6-VL [5]	2025-07-15	**47.56**	**52.30**	**47.76**
Gemini-2.5-Pro [3]	2025-06-05	36.56	42.43	36.41
Qwen-VL-Max [9]	2025-04-08	34.09	40.19	34.12
Open-Source MLLMs				
MiniCPM-V-4.5-8B [14]	2025-08-27	18.31	20.26	17.62
GLM-4.1V-10B [11]	2025-07-02	32.12	42.94	30.85
GLM-4.5V-106B [11]	2025-08-11	41.29	45.63	41.13
Qwen2.5-VL-3B [2]	2025-01-28	30.19	32.35	29.68
Qwen2.5-VL-7B [2]	2025-01-28	27.20	31.63	28.39
Qwen2.5-VL-32B [2]	2025-01-28	24.84	31.88	24.64
Qwen2.5-VL-72B [2]	2025-01-28	40.43	45.21	40.25
InternVL3.5-1B [12]	2025-08-25	25.09	30.45	24.97
InternVL3.5-4B [12]	2025-08-25	29.98	32.98	29.59
InternVL3.5-8B [12]	2025-08-25	31.40	35.08	30.75
InternVL3.5-38B [12]	2025-08-25	39.82	43.18	39.76
InternVL3-78B [15]	2025-04-11	24.93	28.86	25.30
Fine-tuned MLLMs				
Qwen2.5-VL-3B(SFT) [9]	2025-01-28	**43.67**	**46.79**	**43.08**

Table 3. Average Recognition Accuracy (%) on Character-level Benchmark. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	Version	AVG. ACC
Closed-Source MLLMs		
Qwen-VL-Max [9]	2025-04-08	**28.8**
Gemini-2.5-Pro [3]	2025-06-05	18.9
GPT-4o [7]	2025-03-26	13.5
GPT-5 [7]	2025-08-07	14.0
Doubao-Seed-1.6-VL [5]	2025-07-15	23.3
Open-Source MLLMs		
GLM-4.1V-10B [11]	2025-07-02	22.6
GLM-4.5V-106B [11]	2025-08-11	25.3
Qwen2.5-VL-3B [2]	2025-01-28	18.6
Qwen2.5-VL-7B [2]	2025-01-28	23.2
Qwen2.5-VL-32B [2]	2025-01-28	15.8
Qwen2.5-VL-72B [2]	2025-01-28	19.4
InternVL3.5-1B [12]	2025-08-25	5.9
InternVL3.5-4B [12]	2025-08-25	5.3
InternVL3.5-8B [12]	2025-08-25	6.3
InternVL3.5-38B [12]	2025-08-25	19.3
MiniCPM-V-4.5-8B [12]	2025-08-27	21.6
Fine-tuned MLLMs		
InternVL3.5-1B(LoRA) [12]	2025-08-25	**25.4**

Table 4. LLM Post-Correction Average Performance based on Recognition results of Doubao-Seed-1.6-VL [5]. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	Version	AVG		
		AR	CR	1-NED
Closed-Source LLMs				
Gemini-2.5-Pro [3]	2025-06-05	37.75	41.90	37.16
Qwen2.5-Max [9]	2025-03-31	43.68	**51.16**	44.43
Doubao-Seed-1.6 [5]	2025-07-15	41.05	45.75	40.77
Kimi-K2 [10]	2025-07-11	**44.80**	49.66	44.59
GPT-4o [7]	2025-03-26	44.44	49.13	44.43
GPT-5 [7]	2025-08-07	40.28	44.04	39.80
Gemini-2.5-Flash [3]	2025-06-17	43.68	48.45	43.52
Gemini-2.5-Flash-Think [3]	2025-06-17	42.47	46.93	42.06
Open-Source LLMs				
Deepseek-V3-671B [4]	2025-03-24	**46.21**	**51.20**	**46.62**
Deepseek-V3.1-671B [4]	2025-08-21	44.58	49.06	44.63
Deepseek-V3.1-Think-671B [4]	2025-08-21	44.58	49.06	**44.63**
Deepseek-R1-671B [6]	2025-05-28	41.23	44.96	40.78

lead to modifications that compromise the fidelity of the original text, particularly in cases involving rare characters, archaic expressions, or domain-specific terminology. Furthermore, this phenomenon highlights a fundamental tension in knowledge-augmented correction: while external knowledge can provide valuable contextual signals, its indiscriminate application may undermine performance when the input is only a small excerpt of a larger document.

3. Experimental Results

The overall performance of page-level and character-level post-correction, with and without the knowledge base, is reported in Tabs. 2–5. In addition, the page-level and character-level results of Qwen3-VL are provided in Tab. 6 and Tab. 7.

4. An Exploratory Investigation into Knowledge-Based LLM Post-Correction

This section examines the effect of knowledge base size and text segmentation strategy on post-correction outcomes. Understanding these influences not only sheds light on the factors that determine correction effectiveness but also provides guidance for designing future systems that can better leverage external knowledge and handle diverse text structures. we employ the text2vec-base-Chinese [8] model for embedding and the bge-reranker-base [13] model for reranking, enabling us to assess the effects of both knowledge retrieval and ranking strategies on post-correction outcomes. The whole result can be seen from Tab. 8 and Tab. 9.

4.1. Setting

Chunk Strategy1. Specifically, documents shorter than 1,000 characters are kept intact with no overlap; documents between 1,000 and 5,000 characters are chunked with a

Table 5. Knowledge-based LLM Post-Correction Average Performance based on Recognition results of Doubao-Seed-1.6-VL [5]. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	Version	AVG		
		AR	CR	1-NED
Closed-Source LLMs				
Kimi-K2 [10]	2025-07-11	45.44	50.03	45.26
Qwen2.5-Max [9]	2025-01-29	39.64	44.20	39.16
Gemini-2.5-Pro [3]	2025-06-05	40.64	42.93	39.89
Gemini-2.5-Flash [3]	2025-06-17	46.25	**51.69**	46.34
Gemini-2.5-Flash-Think [3]	2025-06-17	**47.10**	50.55	**46.55**
Gemini-2.5-Flash(Reranker) [3]	2025-06-17	46.36	51.82	46.45
Open-Source LLMs				
Deepseek-V3-671B [4]	2025-03-24	46.64	50.10	46.27
Deepseek-V3.1-671B [4]	2025-08-21	42.40	45.81	42.00
Deepseek-V3.1-Think-671B [4]	2025-08-21	**48.67**	**51.03**	**48.05**
Deepseek-R1-617B [6]	2025-05-28	43.22	46.76	42.44

Table 6. The Performance of Qwen3-VL Series on Page-level Recognition Accuracy(%).

Models	AncientBook			JianDu			Calligraphy			Inscription		
	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED
Qwen3-VL-2B [1]	30.67	34.31	30.96	17.53	19.40	17.15	54.49	57.73	53.79	68.62	70.40	68.12
Qwen3-VL-4B [1]	38.07	42.90	38.40	20.49	22.19	19.98	54.84	60.20	53.69	70.83	72.34	70.61
Qwen3-VL-8B [1]	56.00	66.02	57.72	**26.84**	**29.38**	**26.34**	**61.06**	**65.55**	**60.62**	**77.70**	**79.09**	**77.46**
Qwen3-VL-32B [1]	**62.30**	**72.80**	**63.80**	21.25	23.62	20.51	56.58	64.90	56.04	75.18	77.06	75.07
Qwen3-VL-Plus [1]	**61.40**	**70.75**	**62.58**	**26.72**	**28.88**	**26.13**	**64.81**	**69.14**	**64.15**	**78.97**	**80.32**	**78.83**

Table 7. The Performance of Qwen3-VL Series on Character-level Recognition Accuracy(%). The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	ACC (%)					
	Anc.	Jian.	Calli.	Ins.	Silk	Oracle.
Qwen3-VL-2B [1]	65.4	11.2	44.6	**25.2**	4.9	2.4
Qwen3-VL-4B [1]	68.8	13.2	44.0	21.2	4.4	3.2
Qwen3-VL-8B [1]	**72.2**	**17.0**	**57.0**	**29.2**	**5.0**	4.0
Qwen3-VL-32B [1]	66.8	14.2	46.6	18.0	4.0	**4.8**
Qwen3-VL-Plus [1]	**73.8**	**16.2**	**49.4**	22.8	**6.4**	**4.4**

1,000-character window and a 200-character overlap; texts between 5,000 and 10,000 characters are segmented with a 2,000-character window and a 400-character overlap; and texts exceeding 10,000 characters are divided into 4,000-character segments with a 600-character overlap. Introducing progressively larger overlaps for longer texts helps preserve cross-boundary context and reduces information loss, while keeping shorter documents whole preserves their full context and avoids unnecessary fragmentation. This hierarchical, overlap-aware design stabilizes embedding quality, improves semantic coherence across segments, and enhances retrieval performance in downstream knowledge-base construction and RAG tasks.

Chunk Strategy2. Specifically, documents shorter than 1,000 characters are kept intact with no overlap; documents between 1,000 and 5,000 characters are chunked with a 600-character window and an 80-character overlap; texts between 5,000 and 10,000 characters are segmented with an 800-character window and a 120-character overlap; and texts exceeding 10,000 characters are divided into 1,000-character segments with a 200-character overlap.

Raw Knowledge Base: a large knowledge base with 1.7 billion characters constructed from heterogeneous historical texts.

MCHDoc Knowledge Base: a compact knowledge base constructed from the reference text segments corresponding to the evaluation corpus.

4.2. Findings

Impact of Knowledge Base Size on Correction Performance. We observe that the size and composition of the external knowledge base substantially influence the final post-correction performance. Despite the much smaller scale of MCHDoc knowledge base, the curated knowledge base yields significantly larger correction gains. This suggests that relevance and precision of retrieved evidence play a more crucial role than sheer knowledge base size. A large but noisy corpus tends to introduce distractive or weakly related references, whereas a highly targeted knowledge base provides clean and authoritative signals that better guide the correction of OCR outputs. As shown in Tab. 8, On AncientBook, the MCHDoc knowledge base yields an average improvement of 4 percentage points in post-correction accuracy. On Calligraphy, the gain is even more pronounced, reaching 8 percentage points on average. Notably, for JianDu, using the smaller knowledge base not only surpasses the performance obtained with the large corpus but even exceeds the accuracy of the original OCR outputs, indicating that highly relevant and noise-free references are particularly beneficial for this carrier.

Effect of Chunk Strategy on Correction Performance. We further compare two chunking strategies and find that using larger chunks with greater overlap leads to more effective retrieval and larger gains in post-correction performance (Tab. 8 and Tab. 9). Larger and more redundant chunks preserve broader contextual information, which helps the retriever return evidence that is semantically closer to the query. This enriched contextual alignment provides stronger guidance for the correction model, resulting in more accurate revisions of OCR outputs.

Under the MCHDoc-sized knowledge base, however, the difference between the two chunking strategies remains relatively small. The moderate size and domain focus of MCHDoc appear to provide sufficient contextual coverage for retrieval, reducing the sensitivity to how the text is segmented. The only notable exception is observed on the JianDu carrier, where the first strategy—using larger and more overlapping chunks—yields a clear improvement.

In contrast, when applied to the large raw knowledge base, the impact of the chunking strategy becomes sub-

stantially more pronounced. The first strategy consistently outperforms the smaller-chunk alternative, as the preserved contextual breadth helps counteract the noise and heterogeneity inherent in the large corpus. By maintaining more coherent semantic units, the retriever is better able to surface relevant evidence from the noisy background, ultimately yielding stronger retrieval-augmented post-correction performance.

5. Limitation and Future Work

Although our benchmark provides a comprehensive evaluation across multiple carriers, achieving a universal framework for reading Chinese historical documents remains challenging. Additionally, due to copyright restrictions and the limited availability of source materials, the dataset currently does not include page-level images of Oracle Bone and Silk.

Future research can strengthen knowledge-guided post-correction by exploring more advanced retrieval and knowledge integration mechanisms. First, a promising direction is to adopt multi-hop retrieval with deepresearch, where the system recursively expands and verifies evidence rather than relying on a single round of retrieval. Secondly, beyond unstructured retrieval, integrating structured knowledge sources, including classical Chinese knowledge graphs, hierarchical lexical databases, and cross-edition textual alignments, may provide more reliable semantic grounding.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report, 2025. 5

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 5, 6

[4] DeepSeek-AI, Aixin Liu, Bei Feng, et al. Deepseek-v3 technical report, 2025. 5, 6

[5] Dong Guo, Faming Wu, Feida Zhu, et al. Seed1.5-vl technical report, 2025. 5, 6, 8

[6] Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 5, 6

[7] Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. Gpt-4o system card, 2024. 5

[8] Xu Ming. text2vec: A tool for text to vector, 2022. 5

[9] An Yang Qwen, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. 5, 6

[10] Kimi Team, Yifan Bai, Yiping Bao, et al. Kimi k2: Open agentic intelligence, 2025. 5, 6

[11] V Team, Wenyi Hong, Wenmeng Yu, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. 5

[12] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. 5

[13] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023. 5

[14] Tianyu Yu, Zefan Wang, Chongyi Wang, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. 5

[15] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 5

Table 8. Knowledge-based LLM Post-Correction Accuracy with Strategy 1 based on Recognition Results of Doubao-Seed-1.6-VL [5]. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively. Values with **Green** borders denote better than recognition results.

Models	AncientBook			JianDu			CalliGraphy			Inscription		
	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED
Raw knowledge base												
Kimi-K2	47.78	50.51	48.56	18.87	21.55	17.70	59.79	64.69	58.95	55.33	63.36	55.82
Deepseek-V3.1-671B	56.93	61.57	58.57	16.40	19.54	15.16	56.51	61.63	55.40	39.77	40.52	38.88
Deepseek-V3.1-Think-671B	61.33	62.85	61.95	18.95	21.36	17.91	**73.40**	**75.46**	**72.92**	41.00	44.47	39.40
Qwen2.5-Max	33.31	35.38	34.04	17.73	20.56	16.41	53.80	60.17	52.91	53.72	60.70	53.28
Gemini-2.5-Pro	59.92	61.89	60.73	13.95	15.70	12.99	58.70	61.50	57.68	29.98	32.63	28.15
Gemini-2.5-Flash	59.34	64.61	61.36	17.99	20.45	16.77	56.64	62.14	55.71	51.02	59.56	51.51
Gemini-2.5-Flash(Reranker)	59.70	64.93	61.75	18.06	20.72	16.91	56.23	61.87	55.43	51.45	59.75	51.70
Deepseek-V3-671B	58.35	61.48	59.33	17.54	19.70	16.38	63.95	67.03	63.12	46.73	52.19	46.26
Deepseek-R1-671B	60.99	64.16	62.32	15.84	18.13	14.62	61.18	65.49	60.17	34.87	39.25	32.66
Gemini-2.5-Flash-Think	**64.11**	**68.05**	**65.64**	17.21	19.22	15.87	63.02	66.50	62.16	44.07	48.42	42.51
Qwen3-Max	58.79	63.62	60.47	19.13	21.85	18.17	59.46	64.29	58.45	**57.93**	**65.36**	**58.77**
MCHDoc knowledge base												
Kimi-K2	56.34	59.52	57.42	**20.02**	**22.75**	**19.32**	65.33	69.73	64.67	57.15	63.48	57.31
Deepseek-V3.1-671B	58.71	63.10	60.21	18.58	21.69	17.50	63.13	67.54	62.43	41.15	48.02	40.42
Deepseek-V3.1-think-671B	58.36	59.78	58.95	18.82	21.14	17.79	**73.76**	**75.72**	**73.22**	40.84	44.51	39.31
Qwen2.5-Max	34.08	36.07	34.82	18.61	21.78	17.40	58.20	64.30	57.52	53.16	60.02	52.82
Gemini-2.5-Pro	63.95	66.01	64.87	17.56	19.69	16.66	68.55	70.98	67.85	30.83	33.27	29.31
Gemini-2.5-Flash	60.91	65.82	62.87	18.77	21.20	17.66	61.52	66.53	60.79	52.49	59.33	52.70
Gemini-2.5-Flash(Reranker)	60.87	65.77	62.73	18.85	21.34	17.99	61.04	66.31	60.46	52.98	59.54	53.01
Deepseek-V3-671B	59.99	62.84	60.98	19.47	21.59	18.47	71.63	73.96	71.09	47.34	51.21	46.53
Deepseek-R1-671B	63.70	66.59	65.03	18.83	21.39	17.80	70.19	73.59	69.59	36.67	40.30	34.64
Gemini-2.5-Flash-Think	**67.25**	**70.73**	**68.63**	18.65	20.88	17.89	71.89	74.76	71.39	45.90	49.66	44.64
Qwen3-Max	60.24	64.68	61.88	**19.96**	**22.54**	**19.04**	63.31	67.95	62.50	**58.64**	**64.88**	**59.18**

Table 9. Knowledge-based LLM Post-Correction Accuracy with Strategy 2 based on Recognition Results of Doubao-Seed-1.6-VL [5]. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively. Values with **Green** borders denote better than recognition results.

Models	AncientBook			JianDu			CalliGraphy			Inscription		
	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED
Raw knowledge base												
Kimi-K2	52.28	56.00	53.44	17.46	20.12	16.34	50.81	57.48	49.77	**57.07**	**63.89**	**57.25**
Deepseek-V3.1-Think-671B	54.13	56.57	54.81	15.08	17.97	13.79	47.99	54.73	46.76	44.74	55.66	45.28
Qwen2.5-Max	30.79	32.97	31.54	16.39	19.35	14.90	47.09	54.73	45.75	51.26	59.17	50.86
Gemini-2.5-Pro	55.03	57.42	55.93	10.12	11.68	9.05	39.77	44.24	38.06	31.22	34.26	29.63
Gemini-2.5-Flash	57.70	**63.45**	**59.64**	16.60	19.48	15.56	46.87	54.02	45.61	51.19	60.84	51.83
Gemini-2.5-Flash(Reranker)	**57.87**	63.24	59.64	16.32	18.72	15.33	46.11	53.30	45.06	50.56	59.83	51.06
MCHDoc knowledge base												
Kimi-K2	51.02	55.97	51.98	**19.62**	**22.02**	**18.70**	**64.16**	**68.66**	**63.50**	**58.40**	**64.06**	**58.29**
Deepseek-V3.1-Think-617B	58.26	60.69	59.89	18.43	21.15	16.94	62.02	66.60	61.38	42.31	50.10	41.66
Qwen2.5-Max	32.69	34.60	33.37	14.35	16.50	13.37	57.91	63.87	56.97	25.23	28.89	25.21
Gemini-2.5-Pro	**64.53**	**66.94**	**65.57**	17.27	19.25	16.46	**66.16**	**68.98**	**65.47**	31.57	34.05	29.98
Gemini-2.5-Flash	61.46	66.40	63.37	18.64	21.09	17.62	60.00	65.41	59.39	52.13	59.64	52.43
Gemini-2.5-Flash(Reranker)	61.10	66.14	62.95	**18.83**	**21.33**	**17.90**	61.04	66.16	60.43	52.91	59.49	53.06