MCHDoc: A Comprehensive Benchmark for Reading Multi-Carrier Chinese Historical Documents

Yijun Sheng[1,2,†], Shipeng Zhu[1,2,†], Ruijia Zuo[1,2], Na Nie[3,4], Hui Xue[1,2,*]
[1]School of Computer Science and Engineering, Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
[3]Nanjing University Museum, Nanjing University
[4]The China Centre for Linguistic and Strategic Studies, Nanjing University
{220246417,shipengzhu,220252286}@seu.edu.cn , niena@nju.edu.cn , {hxue}@seu.edu.cn

Abstract

Reading Chinese historical documents across diverse carriers is central to understanding the evolution of Chinese civilization, yet remains labor-intensive and dependent on scarce expert knowledge. Although recent large-scale models show promise on isolated historical collections, they do not systematically probe the fundamental ability to read historical documents across heterogeneous carriers. Therefore, we present MCHDoc, a comprehensive benchmark for reading multi-carrier Chinese historical documents. MCHDoc spans over 3,000 years of history and contains 15,724 high-resolution documents from six major carriers, capturing rich variations in material, layout, etc. Mimicking expert workflows, the benchmark supports page-level and character-level recognition, as well as LLM-based post-correction with and without external knowledge. We systematically evaluate a wide range of large-scale models on MCHDoc. The results show that even top-tier models struggle with multi-carrier historical documents. Furthermore, our analysis highlights several key factors for effectively adapting large models to Chinese historical texts. MCHDoc thus offers a standardized, challenging, and historically grounded benchmark for reading Chinese historical documents and provides a foundation for future research in document analysis and digital humanities. The dataset will be released in https://github.com/blackprotoss/MCHDoc.

1. Introduction

Chinese historical documents embody the extensive and profound traditional Chinese culture. Over the long course of history, the carriers of these texts have undergone continuous evolution, ranging from early forms such as Silk to later manuscripts on Paper, each reflecting distinct calligraphic and aesthetic characteristics, as shown in Fig. 1(a). Such artifacts enable digital preservation and support further historical studies [41]. In the past, *experts manually recognized characters from documents and then consulted other professional historical sources or relied on their own domain knowledge to correct these recognized texts*, as shown in Fig. 1(b). However, such traditional methods are time-consuming, labor-intensive [33], and heavily dependent on experts with highly differentiated expertise tailored to each historical carrier [11]. For instance, experts profi-

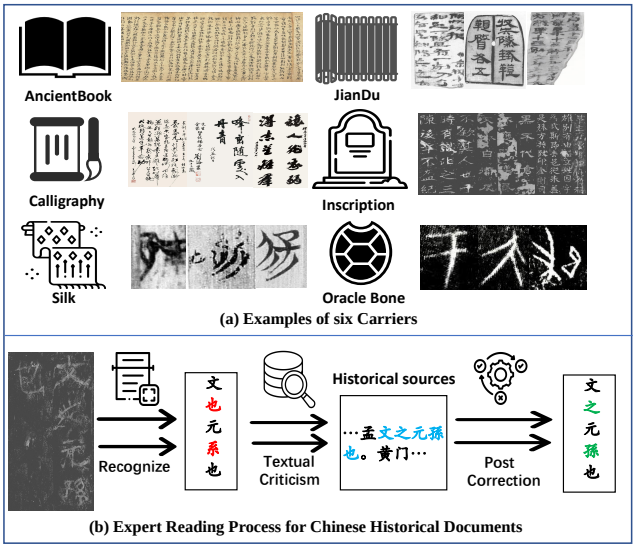


Figure 1. (a) Examples of six carriers. (b) Expert reading process for Chinese historical documents (Red characters: uncertain, Blue text: relevant reference, Green characters: rectified)

[†]Equal contribution, *Corresponding author.

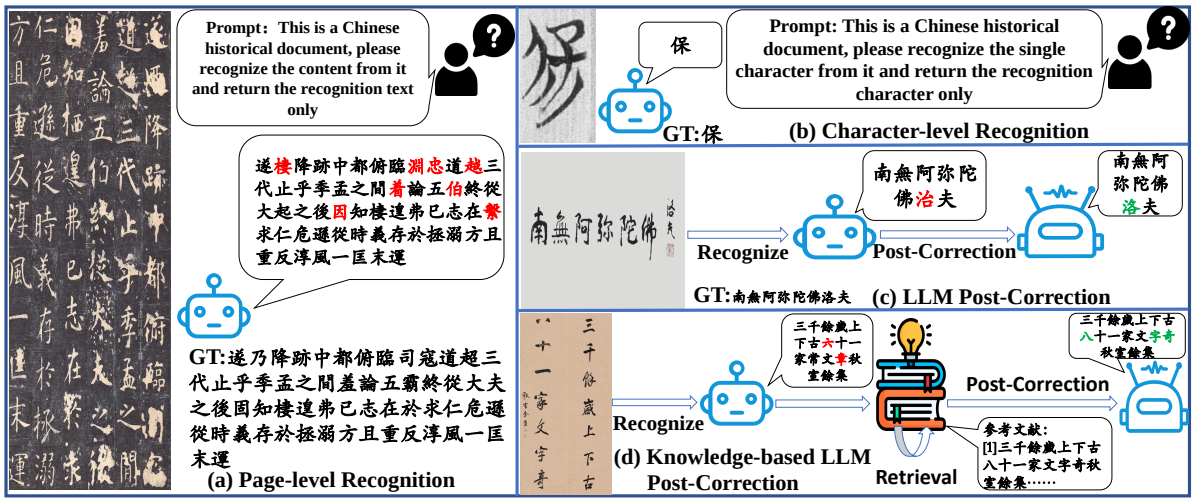


Figure 2. The Illustration of four Evaluation Tasks for Recognition and Post-Correction on Chinese Historical Documents (Red characters: uncertain, Green characters: rectified)

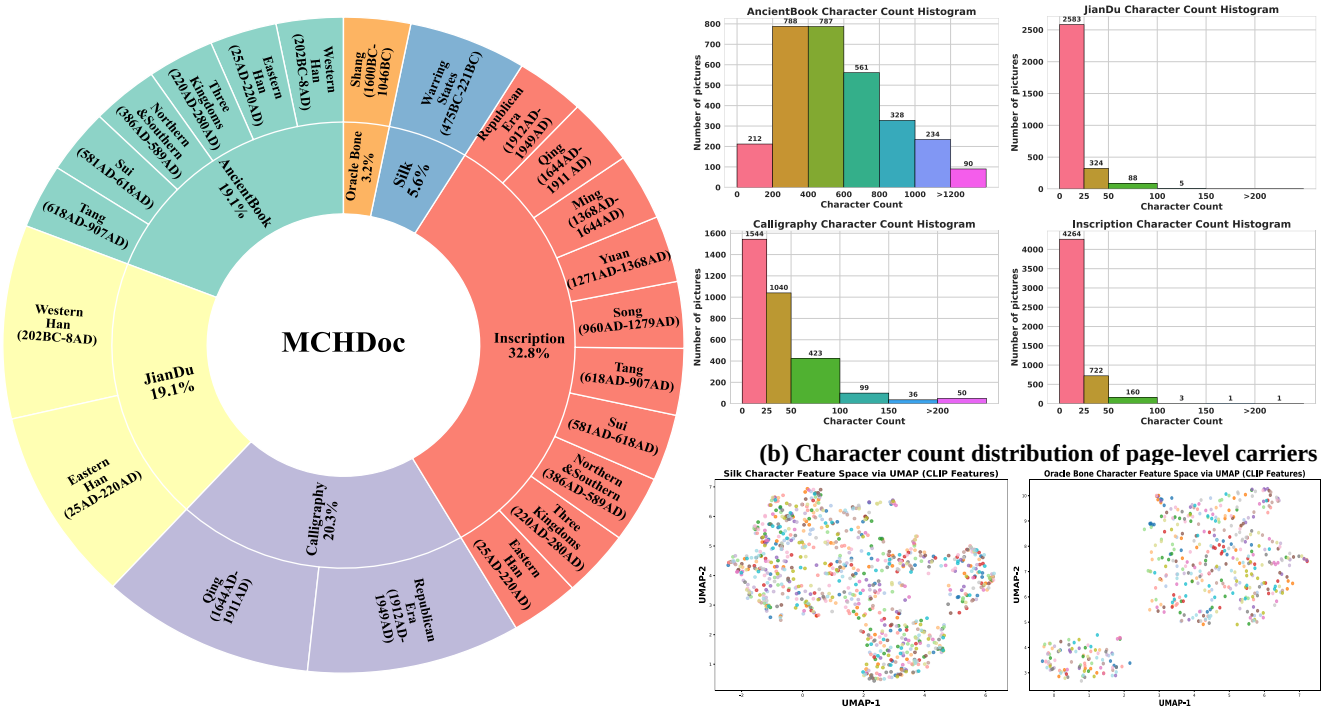
cient in reading books from the Ming dynasty may not be able to transcribe the Oracle Bone from the Shang dynasty precisely.

Recently, Multimodal Large Language Models (MLLMs) and Large Language Models (LLMs) [1, 10, 14] have shown potential in dealing with Chinese documents [12, 34, 37]. As a result, we aim to explore their ability to read Chinese historical documents. Although several studies have conducted preliminary explorations on reading Chinese historical documents [21, 23, 38], several limitations still remain. Firstly, most studies are devoted to Chinese cultural analysis [2, 21, 38], yet paying little attention to the basic reading capability. Secondly, some works focus only on the document recognition [21, 38], ignoring the importance of the text post-correction by verification and citation, which is essential for maintaining the integrity of historical records and enabling reliable scholarly study. Thirdly, the recent Chinese historical documents benchmarks are restricted to a narrow range of carriers [2, 22, 23, 38] and comprise a relatively small number of documents [28, 38, 45], limiting their ability to support broad and systematic studies of Chinese history.

To tackle these challenges, we introduce a comprehensive benchmark for reading Chinese historical documents. It features a wide array of characteristics reflecting both diversity and breadth, containing **over 15,000 annotated images across six main carriers in history**: AncientBook (Hemp Paper), JianDu (Bamboo Slips), Calligraphy (Xuan Paper), Inscription, Silk, and Oracle Bone. Notably, to enhance the diversity and temporal span of carriers, we manually annotated a large number of inscription-based documents. This enriched dataset provides a holistic benchmark for evaluating model performance across various historical contexts on different materials. Our benchmark bridges different dynasties, covering varied visual forms, damage levels, and calligraphic styles that serve as a rigorous testbed for evaluating comprehensive reading capability of large multimodal models. Based on this benchmark, we conduct a comprehensive evaluation of over 20 representative MLLMs and LLMs through four reading tasks for the MCHDoc (Fig. 2), mimicking the workflow of human experts: (1) Page-level recognition, (2) Character-level recognition, (3) Traditional Chinese text post-correction with internal knowledge, (4) Traditional Chinese text post-correction with an external ancient Chinese text knowledge base. The results demonstrate that while top-tier MLLMs and LLMs attain encouraging results on certain carriers, their performance fluctuates significantly across different carriers. This notable variation highlights the challenge of achieving true cross-carrier generalization. The further analysis of results reveal several interesting phenomena that warrant deeper investigation.

In summary, our main contributions are:

- **A large-scale, multi-carrier benchmark** for reading Chinese historical documents, featuring unprecedented diversity across six carrier types ranging from the 16th century BC to the 20th century AD to facilitate robust model evaluation.
- **A thorough evaluation of the reading capabilities of MLLMs and LLMs on MCHDoc**. We conduct a thorough experiment of testing open-source models and closed-source models for reading the MCHDoc.
- **A comprehensive analysis of the results,** revealing their existing strengths and shortcomings, thereby providing directions for Chinese historical culture research.



(a) Distribution of carriers across historical periods

(b) Character count distribution of page-level carriers

(c) Feature distribution of character-level carriers

Figure 3. Overview and Statistics of Multi-carrier Chinese Historical Documents (MCHDoc).

2. Related Work

2.1. Benchmarks for Chinese Historical Documents

In recent years, document reading tasks have made some preliminary progress, supported by the development of deep learning [15, 30] and the availability of large-scale datasets [5, 43]. Within this field, Chinese historical documents present additional challenges due to their complex layouts, diverse scripts, and various types of degradation. A number of datasets have been proposed to facilitate research on these materials. M5HisDoc [28] serves as a large-scale collection of over 8,000 ancient books. HisDoc-1B [29] extends this resource by enlarging the data scale and improving annotation accuracy and scenario diversity. AncientDoc [38] includes more than 3,000 ancient books and defines five representative tasks for evaluation. In addition to textual corpora, several datasets focus on other historical carriers. DeepJianDu [22] contains digitized images of bamboo slip manuscripts with aligned Chinese semantic annotations. CalliBench [23] provides data for Chinese Calligraphy analysis. CIRI [45] focuses on Inscription restoration using images augmented with synthetic noise, and OBI-Bench [2] covers Oracle Bone with multiple tasks such as classification and retrieval.

2.2. Chinese Text Post-Correction

Text correction aims to automatically detect and rectify spelling, grammatical, or semantic errors in text. Early neural approaches [17, 18] formulated Chinese text post-correction as a sequence labeling or sequence-to-sequence problem. PLOME [20] introduces a masking strategy based on a confusion set that replaces masked tokens with visually or phonetically similar characters during pre-training, enabling the BERT model [6] to better understand common error patterns and increase correction accuracy. A training-free LLM-based approach [42] has been developed for general Chinese character error correction, addressing not only traditional substitution errors but also insertion and deletion errors that are often overlooked. GrammarGPT [7] demonstrates that open-source LLMs can achieve strong performance on Chinese grammatical error correction through instruction tuning on a small hybrid dataset with error invariant augmentation. However, existing text correction models developed for modern Chinese are inadequate for ancient Chinese texts, due to significant differences in grammar and linguistic structure.

2.3. Retrieval-Augmented Generation

Querying external knowledge bases requires mechanisms for effectively accessing and leveraging relevant information. Retrieval-Augmented Generation (RAG) [16] in-

tegrates retrieval approaches into large language models to improve factual grounding and contextual reasoning. By retrieving relevant evidence, RAG enables models to overcome the limitations of parametric memory and enhances their robustness in knowledge-intensive scenarios. In the context of reading Chinese historical documents, such retrieval-augmented frameworks can serve as reliable post-correction tools by correcting the original text against real textual sources.

3. MCHDoc

In Section 3.1, we present the sources of different carriers in MCHDoc, followed by a description of the manual labeling process for the Inscription carrier in Section 3.2. Then, In Section 3.3, we provide a statistical analysis of our benchmark and conduct comparisons with previous benchmarks.. Finally, In Section 3.4, we describe how the external knowledge base is constructed.

3.1. Data Collection

Our data collection is distinguished by its comprehensive inclusion of multi-carrier Chinese historical documents, moving beyond the single carrier focus typical of prior studies. The corpus covers six representative carriers: (1) AncientBook (Hemp Paper), sourced from M5HisDoc [28], representing traditional bookmaking widely used throughout imperial China. (2) JianDu (Bamboo Slips) from DeepJianDu [22], excavated primarily in the northwestern regions of China, preserving early administrative and legal records. (3) Calligraphy (Xuan Paper) from CalliBench [23], providing high-quality samples of literary and artistic writing that reflect classical scholarly culture. (4) Inscription, collected from museum holdings, documenting carved historical records on stone or metal surfaces that serve as direct epigraphic evidence. (5) Silk, a key medium for manuscripts and artworks, is adopted from the Wa-net dataset [19] and represented in a character-level format due to its unique degradation and preservation characteristics. (6) Oracle Bone, drawn from OBI-Bench [2] and discovered at the Yinxu archaeological site, provided in character-level format to capture the earliest stages of Chinese writing used in divination practices. By integrating both page-level and character-level data across these materially and geographically diverse carriers, our dataset offers a more holistic foundation for research on multi-carrier Chinese historical documents.

3.2. Inscription Document Annotation

The annotation of inscription documents, particularly rubbings of steles and other carved artifacts, poses unique challenges due to the **complex degradation and stylistic characteristics** of ancient scripts [45]. Each inscription image was transcribed by annotators trained in ancient Chinese epigraphy and subsequently verified through peer review. For ambiguous characters, annotators consulted relevant literature and epigraphic resources, and disagreements were resolved through third-party adjudication. After verification, all inscriptions were systematically organized by dynasty and inscription name, forming a structured corpus for downstream research and cultural heritage preservation.

3.3. Benchmark Analysis

Building on the above efforts, we construct MCHDoc with a total of 15,724 Chinese historical document pages and conduct a detailed analysis of its composition. Concretely, the dataset spans six carrier types from the 16th century BC to the 20th century AD (Fig. 3(a)), including AncientBook (Hemp Paper), JianDu (Bamboo Slips), Calligraphy (Xuan Paper), Inscription (Stone), Silk, and Oracle Bone. Fig. 1(a) shows representative examples of each carrier, highlighting their distinct visual characteristics and material textures. Each carrier contributes a different proportion to the overall corpus, naturally reflecting the diversity and imbalance inherent in real-world historical materials.

Compared with previous benchmarks (Tab. 1), MCHDoc offers a substantially more comprehensive and realistic evaluation setting. It jointly supports both recognition and post-correction tasks, extends the scope from single carrier collections to six materially heterogeneous and historically grounded carriers, and scales the corpus to a larger and more diverse benchmark that is substantially larger than existing Chinese historical document recognition benchmarks. This design better captures the real-world complexity of reading Chinese historical documents.

To provide a deeper understanding of the dataset composition, we further analyze each carrier in terms of image scale, average page resolution, and average number of characters per page. As reported in Tab. 2, image resolutions vary significantly across carriers, from 77×135 pixels for small Oracle Bone fragments to over 2,300×2,000 pixels for AncientBook pages. Moreover, for page-level carriers, as shown in Fig. 3(b), AncientBook pages typically contain hundreds of characters, with a relatively balanced distribution between 200 and 1,000 characters. In contrast, JianDu, Calligraphy, and Inscription images usually contain fewer than 100 characters per image. We also analyze character-level feature distributions for Silk and Oracle Bone by visualizing CLIP-based [27] embeddings with UMAP [24]. As illustrated in Fig. 3(c), characters on Silk form relatively compact clusters, whereas Oracle Bone characters exhibit a much more scattered structure, indicating higher diversity within class and stronger shape variation. This aligns with their different recording carriers and writing forms: AncientBook pages are paragraph-based, while other carriers (e.g., Inscription, JianDu) are dominated by fragmented or short textual content.

Table 1. Comparison between MCHDoc and Previous Chinese Historical Document Benchmark. ✓ indicates support for the task.

Benchmarks	Scale	Carrier						Task	
		AncientBook	JianDu	Calligraphy	Inscription	Silk	Oracle Bone	OCR	Post-Correction
M5HisDoc [28]	8,000	✓	×	×	×	×	×	✓	×
CalliBench [23]	3,192	×	×	✓	×	×	×	✓	×
AncientDoc [38]	3,000	✓	×	×	×	×	×	✓	×
OBI-Bench [2]	1,500	×	×	×	×	×	✓	✓	×
MCHDoc(Ours)	15,724	✓	✓	✓	✓	✓	✓	✓	✓

Table 2. Statistics of Different Carriers

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

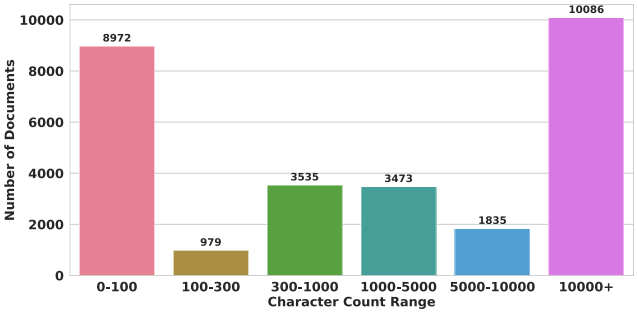


Figure 4. Overview of Knowledge Base Data Distribution.

Overall, these analyses reveal three key challenge factors in MCHDoc: large variation in image pixels, highly diverse and imbalanced character count distributions, and glyph shape diversity across carriers. Together, they make MCHDoc a challenging and comprehensive benchmark for Chinese historical document reading.

3.4. Knowledge Base Construction

Post-Correction with references is the most rigorous reading method. In view of the overwhelming volume of historical sources in real life, the knowledge base is a vast collection of Chinese historical documents, functioning as an essential reference for post-correction. The primary source for our knowledge base is Daizhige [8], which functions as a comprehensive database of ancient Chinese texts. It encompasses nearly 16,000 ancient books across ten major categories, including Classics, Histories, Philosophers, Collections, Poetry and Arts, I Ching studies, Medicine, Buddhism, and Taoism. **The total number of characters in the knowledge base surpasses 1.7 billion**. During the knowledge base construction, we adopt an adaptive, hierarchical chunking strategy to accommodate the substantial variation in document lengths within the corpus (Fig. 4). The detailed chunking strategy is provided in the Appendix.

4. Experiment

4.1. Experiment Protocol

In this benchmark, we evaluate multiple MLLMs/LLMs, including both locally open-source models, fine-tuned models, and remote API-based closed-source models. All the versions were released before 2025 October 1st. In evaluation, some models may fail to return results for certain samples due to various reasons, including model limitations, network issues, or violations of the API security guidelines. All such missing predictions, which account for a small number of cases (overall less than 3%), are treated as failed samples (counted as 0) in metric computation to reflect both model performance and practical usability.

4.2. Evaluation Metric

To thoroughly assess the model performance, we employ three representative metrics: Accurate Rate (AR) [25], Correct Rate (CR) [25], and 1-NED (1-Normalized Edit Distance) [40] to compute the similarity between the predicted and ground-truth strings.

$$\begin{aligned} AR &= (N_t - D_e - S_e - I_e)/N_t, \\ CR &= (N_t - D_e - S_e)/N_t, \end{aligned} \tag{1}$$

Where D_e, S_e, and I_e represent the total number of deletion, substitution, and insertion errors, respectively, and N_t is the total number of characters in the annotations.

$$\text{1-NED} = 1 - \frac{D_{\text{edit}}}{\max(L_{\text{pred}}, L_{\text{true}})}, \tag{2}$$

Where D_{edit} is the Levenshtein edit distance, defined as the minimum number of single-character operations (insertion, deletion, or substitution) required to transform one string into another. L_{pred} and L_{true} denote the lengths of the predicted and ground-truth sequences, respectively. A higher 1-NED value indicates greater sequence similarity and better overall correction performance.

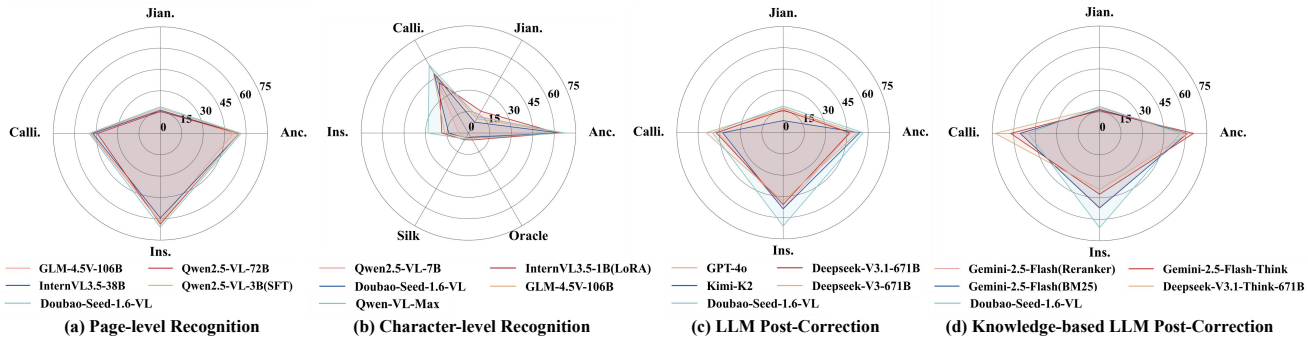


Figure 5. The Performance of Top-Tier Models on four Tasks.

4.3. Results

Page-level Recognition. Page-level recognition is the dominant paradigm in large-scale digitization of Chinese historical documents, so we first evaluate all models in this end-to-end setting [36]. As shown in Tab. 3, the results can be summarized from three perspectives: (1) **Overall capability.** Doubao-Seed-1.6-VL achieves the best overall performance, with Qwen2.5-VL-3B (SFT) and GLM-4.5V-106B as close followers, whereas the GPT series and MiniCPM-V-4.5-8B perform poorly on AncientBook, indicating that generic MLLMs struggle with long-context historical pages. (2) **Carrier-wise difficulty.** Performance varies dramatically across carriers: while top-tier models handle AncientBook and Inscription relatively well, all models perform poorly on JianDu, showing that its severe degradation and unique texture remain challenging. (3) **Scaling and training strategy.** Within the Qwen2.5-VL and InternVL3.5 families, larger variants generally yield better results on harder carriers, and the fine-tuned Qwen2.5-VL-3B (SFT) nearly closes the gap to the closed-source Doubao model, suggesting that domain adaptation is as important as model size.

Character-level Recognition. Character-level recognition offers a more fine-grained view by reducing the reliance on long-range language modeling and focusing on visual decoding. For Silk and Oracle Bone, where page-level samples are scarce, we only conduct character-level experiments. In addition, from the first four carriers, we select pages whose page-level accuracy (1-NED) is below 15%, crop them into individual characters using coordinates, and uniformly sample 500 characters per carrier. As shown in Tab. 4, the main observations are: (1) **Overall capability.** Qwen-VL-Max achieves the best overall performance, with GLM-4.5V-106B and InternVL3.5-1B (LoRA [13]) close behind; Qwen and GLM are consistently strong on AncientBook and Calligraphy, while the GPT series remains clearly inferior. (2) **Carrier-wise difficulty.** All models perform poorly on JianDu, Silk, and Oracle Bone, indicating that even at the character level, their extreme degradation and atypical writing styles are still hard to recognize. (3) **Scaling vs. adaptation.** Parameter scaling does not yield monotonic gains: compact, domain-adapted variants (e.g., InternVL3.5-1B (LoRA [13])) can rival or surpass much larger general models, a trend that echoes but is more pronounced than in the page-level results.

LLM Post-Correction. We further evaluate whether LLMs can improve recognition through post-correction using only their internal knowledge. Concretely, we feed the page-level outputs of Doubao-Seed-1.6-VL [9] into each LLM and let it revise the text without external retrieval. As shown in Tab. 5, we have three main observations: (1) **Overall effect.** Post-correction rarely helps and often hurts: almost all LLMs degrade the original 1-NED, with the inscription carrier suffering the largest performance drop. (2) **Carrier-wise behavior.** For AncientBook, JianDu, and Calligraphy, the changes are generally small, indicating that naive post-correction neither consistently fixes errors nor catastrophically corrupts the outputs. (3) **Reasoning vs. robustness.** Models that perform more aggressive reasoning (e.g., the Gemini and Deepseek series) tend to introduce larger degradations, suggesting that blind LLM-based correction with internal knowledge alone is unsafe for high-stakes historical text digitization.

Knowledge-based LLM Post-Correction. We further evaluate knowledge-based post-correction, where LLMs are augmented with an external epigraphic knowledge base. Complete results are deferred to the Appendix. As shown in Tab. 6, we observe: (1) **Overall gains.** Most models achieve clear improvements on AncientBook and Calligraphy, often surpassing the raw recognition outputs and the internal-only post-correction setting. (2) **Role of reasoning.** Reasoning-enhanced variants (e.g., Gemini-2.5-Flash-Think, Deepseek-V3.1-Think) consistently outperform their non-reasoning counterparts when external knowledge is available, in contrast to the degradation observed with internal knowledge alone. (3) **Failure cases.** On JianDu, the extremely low recognition quality leads to

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

Models	AncientBook			JianDu			Calligraphy			Inscription		
	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED
Closed-Source MLLMs												
GPT-4o [14]	2.41	5.49	3.92	6.59	9.51	4.35	26.78	31.25	25.20	20.00	27.00	18.91
GPT-5 [14]	6.02	7.58	6.87	10.68	11.49	10.42	31.87	34.19	31.23	51.32	52.65	50.86
Doubao-Seed-1.6-VL [9]	**54.05**	**61.62**	**56.52**	**19.52**	**21.68**	**18.82**	**50.71**	**57.64**	**49.81**	**65.96**	**68.27**	**65.90**
Gemini-2.5-Pro [3]	45.98	56.96	48.31	10.59	12.91	8.96	43.79	50.44	42.59	45.86	49.39	45.78
Qwen-VL-Max [26]	30.69	41.12	33.29	14.82	17.51	13.47	47.33	52.99	46.47	43.51	49.13	43.25
Open-Source MLLMs												
MiniCPM-V-4.5-8B [39]	1.44	1.57	1.58	7.80	8.53	7.26	35.25	40.80	33.90	28.76	30.12	27.73
GLM-4.1V-10B [32]	34.26	44.70	35.04	16.14	19.03	15.11	33.90	49.89	30.56	44.18	58.15	42.70
GLM-4.5V-106B [32]	46.37	54.43	47.59	**18.18**	**19.84**	**17.68**	46.45	51.06	45.55	**54.15**	**57.20**	**53.71**
Qwen2.5-VL-3B [1]	30.57	33.89	29.55	7.18	7.68	6.88	33.69	37.70	33.34	49.32	50.13	48.96
Qwen2.5-VL-7B [1]	33.05	38.31	35.46	10.50	11.38	12.63	35.16	40.88	34.54	30.07	36.94	30.93
Qwen2.5-VL-32B [1]	13.91	23.62	17.01	8.86	11.22	7.14	39.63	48.14	38.22	36.98	44.54	36.18
Qwen2.5-VL-72B [1]	44.86	53.13	46.95	17.50	19.35	16.06	48.11	53.68	47.22	51.24	54.68	50.77
InternVL3.5-1B [35]	30.76	39.57	33.10	4.40	5.01	3.78	34.69	41.94	33.88	30.49	35.29	29.13
InternVL3.5-4B [35]	37.45	41.16	38.15	4.25	4.96	3.68	44.75	50.67	43.91	33.47	35.14	32.60
InternVL3.5-8B [35]	42.21	47.54	43.30	5.37	6.67	4.13	43.23	49.42	42.02	34.80	36.68	33.54
InternVL3.5-38B [35]	**52.40**	**57.49**	**54.03**	9.79	11.02	8.57	42.65	47.73	42.65	54.45	56.49	53.79
InternVL3-78B [44]	11.04	16.33	14.18	9.17	10.40	8.39	34.21	40.25	33.99	45.28	48.48	44.63
Fine-tuned MLLMs												
Qwen2.5-VL-3B(SFT) [26]	52.53	56.46	52.53	17.80	21.42	17.09	**55.09**	**57.71**	**54.26**	49.26	51.56	48.44

Table 4. Recognition Accuracy (%) on Character-level Benchmark. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	ACC (%)					
	Anc.	Jian.	Calli.	Ins.	Silk	Oracle.
Closed-Source MLLMs						
Qwen-VL-Max [26]	**67.6**	13.0	**54.8**	**28.2**	4.9	4.0
Gemini-2.5-Pro [3]	43.0	6.4	39.2	15.6	3.2	5.6
GPT-4o [14]	33.0	4.2	29.6	9.8	2.0	2.6
GPT-5 [14]	35.0	5.4	26.6	8.6	3.0	**6.0**
Doubao-Seed-1.6-VL [9]	64.0	8.8	48.0	13.8	3.9	3.4
Open-Source MLLMs						
GLM-4.1V-10B [32]	64.8	8.6	46.8	10.6	2.5	2.4
GLM-4.5V-106B [32]	**66.2**	11.6	**49.0**	16.6	3.8	4.4
Qwen2.5-VL-3B [26]	62.2	8.6	34.4	1.2	2.8	2.4
Qwen2.5-VL-7B [26]	65.6	**13.4**	46.8	7.0	4.0	2.8
Qwen2.5-VL-32B [26]	50.0	6.4	32.0	2.2	2.3	1.8
Qwen2.5-VL-72B [26]	62.0	12.8	34.8	6.8	2.7	2.8
InternVL3.5-1B [35]	13.6	0.3	14.4	5.8	0.8	0.2
InternVL3.5-4B [35]	5.4	1.5	10.5	2.5	1.0	0.4
InternVL3.5-8B [35]	19.6	0.6	12.0	2.2	1.4	0.4
InternVL3.5-38B [35]	35.8	4.4	45.4	**20.0**	5.2	5.0
MiniCPM-V-4.5-8B [35]	39.0	7.6	32.0	11.4	**12.4**	**6.0**
Fine-tuned MLLMs						
InternVL3.5-1B(LoRA) [35]	62.6	**17.6**	40.8	18.6	**5.5**	5.6

unreliable retrieval, leaving little room for effective correction; on Inscription, each cropped image covers only a small fragment of a long stele, so fragment-level queries often miss the correct entry in the knowledge base, resulting in limited gains.

4.4. Discussion

In this section, we conduct a comprehensive analysis of the performance of the model on reading multi-carrier Chinese historical documents and find several factors affecting the ability of the model. More detailed analyses are provided in the **Appendix**.

Importance of Training Corpus. As shown in Fig. 5(a) and (b), models trained with Chinese historical document image recognition generally outperform those without such training. For instance, Doubao-Seed-1.6-VL [9] reported utilizing extensive Chinese historical document images and corresponding text as training data during pre-training, enabling it to achieve the strongest overall performance in multi-source classical text document recognition. To further validate this observation, we constructed a page-level and character-level training dataset (**Appendix**) for multi-carrier Chinese historical text recognition. Qwen2.5-VL-3B [1] significantly outperformed many large-parameter models under full-parameter fine-tuning in page-level recognition, with its overall performance trailing only Doubao-Seed-1.6-VL [1]. InternVL3.5-1B [35] with LoRA [13] also surpasses various large-scale models in character-level recognition.

Limitation of Internal Knowledge. As shown in Fig. 5(c), when leveraging their internal knowledge without external retrieval, LLMs fail to perform effective correction across all six historical carriers. Even for specific carriers such as AncientBook and Inscription, where the recognition ac-

Table 5. LLM Post-Correction Performance based on Recognition Results of Doubao-Seed-1.6-VL [9]. The **Red** and **Blue** denote the optimal and sub-optimal results, respectively.

Models	AncientBook			JianDu			Calligraphy			Inscription		
	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED	AR	CR	1-NED
Closed-Source LLMs												
Gemini-2.5-Pro [3]	47.65	52.43	48.57	13.35	15.22	12.32	43.88	49.81	42.60	46.12	50.15	45.13
Qwen2.5-Max [26]	51.37	60.21	54.26	17.47	**20.98**	16.41	48.12	57.21	47.93	57.76	66.23	59.11
Doubao-Seed-1.6 [9]	51.38	57.82	52.98	16.10	18.14	15.08	41.63	47.67	40.30	55.09	59.35	54.73
Kimi-K2 [31]	**53.38**	**60.50**	**55.39**	17.38	19.32	16.24	48.73	55.34	47.25	59.72	63.46	59.46
GPT-4o [14]	48.95	55.31	50.96	**17.73**	19.96	**16.87**	**49.28**	**56.19**	**48.23**	61.81	65.08	61.65
GPT-5 [14]	48.20	53.14	49.39	15.37	17.43	14.22	46.23	52.95	44.72	51.32	52.65	50.86
Gemini-2.5-Flash [3]	**53.43**	**60.47**	**55.49**	16.78	18.76	15.83	46.13	52.93	44.75	58.39	61.64	58.00
Gemini-2.5-Flash-Think [3]	50.59	56.48	51.86	16.41	18.55	15.48	46.54	52.99	45.10	56.32	59.71	55.81
Open-Source LLMs												
Deepseek-V3-671B [4]	53.31	60.43	55.39	**18.76**	**20.82**	**17.89**	**49.94**	**56.99**	**48.93**	**62.84**	**66.58**	**62.83**
Deepseek-V3.1-671B [4]	52.92	60.35	55.11	16.32	18.19	15.46	45.09	50.84	44.15	**63.99**	**66.87**	**63.80**
Deepseek-V3.1-Think-671B [4]	49.31	54.05	50.73	17.26	19.19	16.35	44.20	49.78	43.20	57.02	60.11	56.60
Deepseek-R1-671B [10]	48.71	52.82	49.76	16.11	17.99	15.17	46.48	52.32	45.11	53.61	56.73	53.09

Table 6. Knowledge-based LLM Post-Correction Performance based on Recognition Results of Doubao-Seed-1.6-VL [9]. 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
Closed-Source LLMs												
Kimi-K2 [31]	47.78	50.51	48.56	**18.87**	**21.55**	**17.70**	59.79	64.69	58.95	**55.33**	**63.36**	**55.82**
Qwen2.5-Max [26]	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 [3]	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 [3]	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-Think [3]	**64.11**	**68.05**	**65.64**	17.21	19.22	15.87	63.02	66.50	62.16	44.07	48.42	42.51
Gemini-2.5-Flash(Reranker) [3]	59.70	**64.93**	61.75	18.06	20.72	16.91	56.23	61.87	55.43	51.45	59.75	51.70
Open-Source LLMs												
Deepseek-V3-671B [4]	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-V3.1-671B [4]	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 [4]	61.33	62.85	61.95	**18.95**	**21.36**	**17.91**	**73.40**	**75.46**	**72.92**	41.00	44.47	39.40
Deepseek-R1-617B [10]	**60.99**	64.16	**62.32**	15.84	18.13	14.62	61.18	65.49	60.17	34.87	39.25	32.66

curacy is relatively high, the models still struggle to revise misrecognized characters once erroneous text is provided as input. For more challenging carriers such as JianDu, the situation is even worse, as the severely noisy text input impairs the capacity of LLMs. These results suggest that self-contained knowledge in LLMs is insufficient for robust correction, emphasizing the necessity of knowledge-grounded approaches.

Effectiveness of External Knowledge Base. As shown in Fig. 5(c) and (d), models that rely solely on the parametric knowledge encoded during pre-training struggle to accurately correct recognition errors, particularly when facing ambiguous glyphs or domain-specific linguistic patterns in Chinese historical documents. In contrast, incorporating external knowledge, such as authoritative text references, provides concrete signals that guide the correction process. With this additional grounding, several models exhibit substantial improvements and are able to produce accurate corrections on carriers such as AncientBook and Calligraphy, underscoring the critical role of external knowledge in enabling robust correction in specialized domains.

5. Conclusion

In this paper, we introduce MCHDoc, a comprehensive and challenging benchmark that assesses the reading capabilities of various large-parameter models on Chinese historical documents across multiple carriers. By covering a wide array of carriers and fine-grained tasks, MCHDoc fills a gap in existing works, providing a holistic evaluation framework for recognition and post-correction tasks. The results show that even top-tier models struggle with multi-carrier Chinese historical documents, and we further find that external historical knowledge base are crucial for post-correction, providing invaluable insights for digital humanities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62476056, T24B2005), the National Social Science Fund of China (No. 25BTQ023), and the Fundamental Research Funds for the Central Universities (2242025K30024). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report, 2025. 2, 7

[2] Zijian Chen, tingzhu chen, Wenjun Zhang, et al. OBI-bench: Can lmms aid in study of ancient script on oracle bones? In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 4, 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. 7, 8

[4] DeepSeek-AI, Aixin Liu, Bei Feng, et al. Deepseek-v3 technical report, 2025. 8

[5] Chao Deng, Jiale Yuan, Pi Bu, et al. LongDocURL: A comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1135–1159, 2025. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 3

[7] Yaxin Fan, Feng Jiang, Peifeng Li, et al. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *CCF international conference on natural language processing and Chinese computing*, pages 69–80. Springer, 2023. 3

[8] Garychowcmu. Daizhigev20. https://github.com/garychowcmu/daizhigev20, 2019. 5

[9] Dong Guo, Faming Wu, Feida Zhu, et al. Seed1.5-vl technical report, 2025. 6, 7, 8

[10] Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 2, 8

[11] Wei Han. Interview 049. In *Daily Progress: Interviews with Young Scholars on Excavated Documents and Paleography*. Center for Research on Chinese Excavated Classics and Paleography, Fudan University, 2020. 1

[12] Anwen Hu, Haiyang Xu, Liang Zhang, et al. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding, 2024. 2

[13] Edward J Hu, yelong shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 6, 7

[14] Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. Gpt-4o system card, 2024. 2, 7, 8

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, et al. Ocr-free document understanding transformer, 2022. 3

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474, 2020. 3

[17] Si Li, Jianbo Zhao, Guirong Shi, et al. Chinese grammatical error correction based on convolutional sequence to sequence model. *IEEE Access*, 7:72905–72913, 2019. 3

[18] Shuangyin Li, Jinbin Zhang, and Yuncheng Jiang. An end-to-end method for chinese spelling error detection and correction. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, pages 232–244, 2024. 3

[19] Shengnan Li, Chi Zhou, and Kaili Wang. Wa-net: Wavelet integrated attention network for silk and bamboo character recognition. *Engineering Applications of Artificial Intelligence*, 140:109674, 2025. 4

[20] Shulin Liu, Tao Yang, Tianchi Yue, et al. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 2991–3000, 2021. 3

[21] Yang Liu, Jiahuan Cao, Hiuyi Cheng, et al. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 10435–10492, 2025. 2

[22] Yiran Liu, Qiang Zhang, Ying Qi, et al. Deepjiandu dataset for character detection and recognition on jiandu manuscript. *Scientific Data*, 12(1):398, 2025. 2, 3, 4

[23] Yuxuan Luo, Jiaqi Tang, Chenyi Huang, et al. Callireader: Contextualizing chinese calligraphy via an embedding-aligned vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23030–23040, 2025. 2, 3, 4, 5

[24] Leland McInnes, John Healy, and James Melville. Umap: uniform manifold approximation and projection for dimension reduction, 2020. 4

[25] Dezhi Peng, Lianwen Jin, Weihong Ma, et al. Recognition of handwritten chinese text by segmentation: A segment-annotation-free approach. *IEEE Transactions on Multimedia*, 25:2368–2381, 2023. 5

[26] An Yang Qwen, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. 7, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 4

[28] Yongxin Shi, Chongyu Liu, Peng, et al. M5hisdoc: A large-scale multi-style chinese historical document analysis benchmark. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2023. 2, 3, 4, 5

[29] Yongxin Shi, Dezhi Peng, Yuyi Zhang, et al. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 12(1):169, 2025. 3

[30] Zineng Tang, Ziyi Yang, Guoxin Wang, et al. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023. 3

[31] Kimi Team, Yifan Bai, Yiping Bao, et al. Kimi k2: Open agentic intelligence, 2025. 8

[32] 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. 7

[33] Xiuan Wan, Yuchun Fang, Jiahua Wu, et al. Geometrics assisted rubbing generation and semantics enhanced detection for small and dense obi character. *Int. J. Interact. Multim. Artif. Intell.*, 9:78–91, 2025. 1

[34] Dongsheng Wang, Natraj Raman, Mathieu Sibue, et al. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8529–8548, 2024. 2

[35] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. 7

[36] Hailin Yang, Lianwen Jin, and Jifeng Sun. Recognition of chinese text in historical documents with page-level annotations. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 199–204. IEEE, 2018. 6

[37] Jiabo Ye, Anwen Hu, Haiyang Xu, et al. UReader: universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*, pages 2841–2858, 2023. 2

[38] Haiyang Yu, Yuchuan Wu, Fan Shi, et al. Benchmarking vision-language models on chinese ancient documents: From ocr to knowledge reasoning. *arXiv preprint arXiv:2509.09731*, 2025. 2, 3, 5

[39] Tianyu Yu, Zefan Wang, Chongyi Wang, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. 7

[40] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1577–1581, 2019. 5

[41] Ruili Zhang, Yanming Yang, and Wenxiu Wang. Research on document digitization processing technology. In *MATEC Web of Conferences*, page 02014. EDP Sciences, 2020. 1

[42] Houquan Zhou, Bo Zhang, Zhenghua Li, et al. A training-free llm-based approach to general chinese character error correction. *arXiv preprint arXiv:2502.15266*, 2025. 3

[43] Fengbin Zhu, Wenqiang Lei, Fuli Feng, et al. Towards complex document understanding by discrete reasoning. In *Proceedings of the ACM International Conference on Multimedia*, pages 4857–4866, 2022. 3

[44] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 7

[45] Shipeng Zhu, Hui Xue, Na Nie, et al. Reproducing the past: A dataset for benchmarking inscription restoration. In *Proceedings of the ACM International Conference on Multimedia*, page 7714–7723, 2024. 2, 3, 4