SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

A Divergence-Driven Framework for Chinese Literary Text Digitization: Multi-Engine OCR Accuracy Evaluation and Multi-LLM Post-Correction

Xin Zhang [1,*], Rachel Suet Kay Chan [1,*] and Raan-Hann Tan [2]

[1] Institute of Ethnic Studies (KITA), Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia

[2] Institute of Malaysian and International Studies (IKMAS), Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia

[*] Correspondence: zx20230918@gmail.com (X.Z.); rachelchansuetkay@ukm.edu.my (R.S.K.C.)

Received: 4 May 2026; **Revised:** 26 June 2026; **Accepted:** 6 July 2026; **Published:** 21 September 2026

Abstract: The digitization of Chinese literary texts for corpus construction faces three challenges: vertical typesetting, a set covering more than 80,000 Chinese characters codified in China's national standard GB 18030, and frequent use of rare characters. To address gaps in existing research on multi-engine comparison and accuracy quantification, this study adopts a three-phase framework: engine evaluation, workflow design, and end-to-end validation. Engine evaluation comprises two experiments. Experiment 1, on 13 pages of vertical Traditional Chinese, compares three engines: Kandian Guji v2.0.7 (Kandian) achieves 0.25% Character Error Rate (CER), far surpassing PaddleOCR-VL-1.5 (Paddle1.5) at 4.77% and ABBYY FineReader 15 (ABBYY) at 5.96%. Experiment 2 extends to four engines on 21 pages of horizontal Simplified Chinese by adding PP-OCRv5 (PP5); Paddle1.5 achieves the lowest CER at 0.74%, and its error distribution diverges across all three error classes from Kandian's, supporting a multi-engine scheme. ABBYY and PP5 are excluded from the final workflow due to error patterns. The four-stage workflow is adopted for the Simplified Chinese scenario: multi-engine Optical Character Recognition (OCR), character-level diff detection, multi-Large Language Model (LLM) arbitration, and final human proofreading. Experiment 3 validates the end-to-end workflow on 20 pages of Simplified Chinese text, where Kandian and Paddle1.5 produce initial CERs of 0.54% and 0.28%; diff detection identifies 62 divergences, of which 49 reach consensus, reducing CER to 0.26%; human proofreading reduces it to 0.00%. For the Traditional Chinese scenario, a rule-based script grounded in character co-occurrence legitimacy is adopted, reducing CER from 0.286% to 0.074%. The framework offers a reproducible quality-control method for Chinese literary digitization.

Keywords: Chinese Literary Corpus; Multi-Engine OCR; LLM Post-Correction; CER; Character-Level Diff Detection; OCR Digitization Workflow

1. Introduction

Converting printed text into computable digital corpora is a prerequisite for large-scale empirical research in corpus linguistics, digital humanities, and computational translation studies [1]. For Chinese literary texts, however, this digitization process confronts several technical challenges. This study makes four contributions. First, it applies multi-engine character-level divergence detection to locate candidate errors. Second, it uses multi-LLM binary arbitration as a routing mechanism between auto-resolved and human-reviewed positions. Third, it reports

end-to-end per-stage CER accounting. Fourth, it defines a separate, auditable rule-based path for vertical Traditional Chinese, where divergence detection is invalid.

First, literary publications from Taiwan and early mainland China predominantly adopt vertical typesetting, with composite page designs that mix horizontal headings and vertical body text. Such multi-directional layouts pose a direct challenge to OCR layout analysis: sorting OCR results from top to bottom and left to right according to the absolute coordinates of detected text boxes yields an order that is usually unstable and inconsistent with the reading order [2]. This limitation is acute on pages combining vertical and multi-column layouts, where column-sequence errors are highly likely.

Second, commercial OCR software struggles with the scale of the Chinese character set. China's mandatory national standard GB 18030-2022 codifies 87,887 Chinese characters [3]; Unicode 17.0 encodes more than 100,000 Chinese, Japanese, and Korean (CJK) Unified Ideographs, with the Ideographic Research Group continuing to submit proposals for new additions [4].

Third, in Chinese literary texts, the rare traditional characters, dialect words, and literary archaisms intentionally used by authors—such as 鑾湧, 曫, 焱, and 朘—constitute core elements of stylistic identity. Their forms are difficult to distinguish from characters introduced by OCR errors; furthermore, the number of visually similar character pairs in Chinese far exceeds that of alphabetic scripts, and visual misrecognitions leave no detectable anomaly at the character level, making automatic error identification nearly impossible. Automatic correction methods that rely on linguistic-statistical priors are therefore prone to misjudgment. This is a challenge that distinguishes the digitization of Chinese literary texts from that of standard printed texts and even literary texts in other languages.

These challenges collectively indicate that high-quality digitization of Chinese literary texts requires not only multi-engine collaboration to compensate for individual OCR failures, but also a post-correction mechanism capable of distinguishing genuine OCR errors from authorial stylistic choices—a distinction that statistical approaches cannot reliably make. Meanwhile, two gaps remain in the existing literature. First, OCR quality evaluation studies have focused predominantly on Western historical-document digitization, including large-scale historic newspaper digitization programs and the assessment of OCR evaluation tools and metrics for mass-digitized historical documents [5, 6]. Multi-engine comparative evaluation for modern literary texts, particularly vertical Traditional Chinese, remains rare. Second, while research on LLM post-correction is growing, end-to-end quantification of the integrated OCR-to-correction workflow remains scarce [7], and traceable per-stage accuracy contributions are largely absent from existing studies.

Addressing these gaps, this paper raises two research questions. Research question 1: For Chinese literary texts, what are the recognition accuracies of mainstream OCR engines on vertical Traditional Chinese and horizontal Simplified Chinese texts, and what differences exist in error-type distributions across engines? Research question 2: Can an end-to-end workflow driven by multi-engine divergence comparison and multi-LLM post-correction reduce initial OCR error rates to a level that meets corpus construction requirements, and what is the quantifiable contribution of each processing stage? The empirical findings from both questions directly inform engine selection and workflow configuration decisions for Chinese literary text digitization. Three experiments are conducted sequentially, using texts drawn from an ongoing Chinese literary translation corpus project. Experiments 1 and 2 evaluate candidate OCR engines to inform workflow configuration; Experiment 3 comprises Sub-Experiment A, which executes the full four-stage workflow on simplified text, and Sub-Experiment B, which documents why the same workflow is inapplicable to traditional text and proposes an alternative correction path. Sample sizes are set by controlled-experiment principles to enable character-level traceability; full-corpus accuracy estimation lies beyond the scope of this paper.

2. Related Work

2.1. Chinese OCR: Technical Evolution and Research Gaps

OCR technology has shifted from template matching to deep learning. Early systems used a three-stage workflow: preprocessing, layout analysis, and feature classification. Pattern matching was the core method. Each module needed separate design. With the rise of deep learning, Convolutional Neural Networks (CNNs) have been widely used for textual feature extraction, and combined with sequence modeling and attention mechanisms they form an end-to-end OCR workflow that breaks the three-stage framework, yielding substantial gains in efficiency and ac-

curacy [8, 9]. Li et al. further showed that initializing both encoder and decoder with pre-trained Transformer models—bypassing CNN backbones entirely—yields state-of-the-art accuracy on printed, handwritten, and scene text tasks [10].

Beyond this, general technical progress does not automatically resolve the special challenges in the Chinese literary scenario. At the character-set level, the CJK character set remains an open and expanding standard, outpacing the coverage capacity of most commercial OCR systems. At the layout level, the complexity of literary texts further increases OCR difficulty: an evaluation of late Qing and Republican vertical Traditional Chinese literary texts reported that four mainstream commercial OCR packages achieved OCR accuracies between only 78.9% and 88.9%—far below the 97% industry benchmark for modern printed texts—pointing to inadequate adaptation to literary texts and calling for purpose-built solutions [11]. A recent study by Liu et al. proposed a multi-oriented detector with dual-path text recognition for Chinese historical documents, achieving end-to-end character-level accuracy competitive with commercial systems on traditional-script corpora [12]. A separate study of one hundred digitized issues of Chao Foon, a Chinese-language Malayan literary journal printed in Traditional Chinese, likewise documented persistent column-order errors in multi-column layouts and recurring missed OCR of variant characters, ultimately requiring human proofreading to correct [13].

2.2. Multi-Engine OCR and LLM Post-Correction

Cao's evaluation first introduced multi-engine comparison in the horizontal Chinese literary scenario [11]. The study stopped at single-engine selection. It identified the best of four engines for independent use. It did not explore cross-validation of multi-engine outputs or divergence-driven quality control. Correction relied entirely on human proofreading without automated assistance. Zavorin et al. demonstrated through the MEMOE system on Arabic document corpora that a multi-engine cascade outperforms any individual engine in OCR accuracy, validating the feasibility of evidence-fusion frameworks and providing engineering precedent for the multi-engine comparison design adopted in this study [14]. Nguyen et al., in their survey of post-OCR correction methods, identified the merging of multiple OCR outputs as a viable direction for quality improvement and noted that extending post-OCR research to non-Latin scripts such as Arabic and Chinese remains under-explored—a gap that directly motivates the present study [15]. Ramirez-Orta et al. demonstrated on the ICDAR2019 multilingual benchmark that large ensembles of character-level sequence-to-sequence models outperform single-model baselines across five languages, providing engineering evidence for multi-model output merging [16]. Rijhwani et al. showed that semi-supervised learning with lexically aware decoding improves post-OCR correction in low-resource settings where annotated data is scarce, highlighting the difficulty of extending post-correction methods to non-standard scripts [17].

For LLM post-correction, existing research provides strong proof-of-concept results. Thomas et al. instruction-tuned Llama 2 on 19th-century English historical-newspaper corpora and achieved a 54.51% CER reduction relative to a BART baseline, demonstrating that generative LLMs offer significant advantages on post-correction tasks even with limited training data [18]. Greif et al. extended this approach to the multimodal setting, showing that a multimodal LLM can achieve 0.84% CER in historical-document post-correction without any image preprocessing or model fine-tuning; the authors further noted that LLM performance on non-Latin-script historical documents remains unknown, identifying this as an important direction for future research [7]. However, Kanerva et al. demonstrated that zero-shot post-correction with open-weight LLMs yields markedly inconsistent results across languages: while CER improvements of up to 39.9% were achieved on historical English texts, the approach failed entirely for historical Finnish [19]. This reveals limitations in the cross-lingual generalizability of single-LLM post-correction and further supports the multi-LLM proofreading design adopted in the present study. The studies above all focus on Latin-script historical documents; their applicability to the Chinese literary scenario remains unverified, and post-correction stages are typically reported in isolation, lacking unified, end-to-end quantitative data on per-stage accuracy contributions.

At the level of evaluation methodology, full-text ground-truth assessment is neither feasible nor affordable in a large-scale digitization context. In their survey of OCR evaluation methods, Neudecker et al. explicitly state that small-sample evaluation is the field's prevailing compromise, and emphasize that without incorporating layout-analysis accuracy, current metrics have limited applicability to downstream tasks such as NLP and digital humanities [6]. The practice of Kettunen et al. provides direct quantitative grounding for the small-sample strategy: in evaluating a historical-newspaper corpus of approximately 16.51 million pages, only 479 pages were randomly

drawn as ground-truth samples, a sampling rate of about 0.003% [20]. As a quality benchmark, Holley defines good OCR accuracy for printed text as a CER of 1%–2% [5]. This benchmark derives from English newspapers; because Chinese is logographic, one character often carries a whole word's lexical load, so the same CER implies greater semantic loss than in English. Therefore 1%–2% is treated as a conservative external reference, consistent with the lower OCR accuracies reported for vertical Traditional Chinese literary texts [11].

The studies above reveal an unaddressed evaluation gap: existing research generally lacks designs that compare multiple engine outputs, and post-correction relies on human proofreading without LLM-based automation; end-to-end evaluation for the modern Chinese literary scenario—coexisting Traditional and Simplified Chinese, with variable scan quality—remains absent. The present paper designs experiments to address these three gaps simultaneously, offering an end-to-end scheme of multi-engine OCR comparison plus multi-LLM post-correction as a reproducible quality-control framework for Chinese literary text digitization.

3. Evaluation of OCR Engines

3.1. Evaluation Design

This section reports engine-comparison evidence: it evaluates candidate OCR engines and supports the engine-selection decisions for the workflow. The two sets of experimental texts reported in this section both come from an ongoing Chinese literary translation corpus project: Experiment 1 targets the project's core Traditional Chinese texts, and Experiment 2 targets the Simplified Chinese texts. The engine-selection conclusions will directly inform that project's digitization workflow configuration. To this end, the present study conducts engine evaluation prior to workflow design, with two purposes: first, to provide a quantifiable comparison of candidate engines' OCR accuracy under uniform controlled conditions; second, to base engine selection on empirical evidence rather than prior assumptions. Evaluation dimensions include CER, inter-engine output consistency, and adaptability to specific layout types. Experimental ground truth (hereafter GT) is established through character-by-character human proofreading using dedicated proofreading tools.

Candidate engines were chosen to vary along three dimensions—architectural paradigm, framework provenance, and licensing model—so that each engine plays a non-redundant role in the comparison. Kandian Guji v2.0.7 (hereafter Kandian) is an OCR platform purpose-built for Traditional and Simplified Chinese documents; its engine includes a dedicated vertical-layout analysis module that restores right-to-left, top-to-bottom reading order for vertical text—an adaptation that general OCR engines commonly lack when handling vertical Traditional Chinese literary texts. Kandian's proofreading interface presents a row-by-row, three-layer stacked structure—original scan image, engine-OCR text, and an editable text box arranged in vertical alignment—allowing the proofreader to verify the image and adjudicate the text within a single view. This dual capacity makes Kandian both a candidate engine and a GT human-proofreading platform (see **Figure 1**). To prevent circularity, ground-truth proofreading was performed blind to the cross-engine comparison. Each character was corrected against the source page image only, never against another engine's output. A single proofreader carried out the work, and uncertain characters were resolved by re-inspecting the source image. The ground truth is therefore image-derived, not Kandian-derived: Kandian supplies a starting draft, but every character is verified against the image, so Kandian serves as a candidate engine and as the proofreading starting point, not as the standard itself. Paddle1.5's independent re-OCR against the same ground truth provides a further cross-check. Before fixing Kandian as the GT-construction tool for Traditional Chinese texts, the study also evaluated two similar OCR engines, namely Shidian Guji and Huicong Guji. The former's output contains numerous candidate-character annotation symbols, making it unsuitable for accuracy evaluation of modern printed literary texts; the latter's overall CER is too high. As a result, neither was included in the engine configuration. PaddleOCR-VL-1.5 (hereafter Paddle1.5) adopts a two-stage "visual understanding and language decoding" architecture: a layout-detection module first identifies page structure, a visual encoder converts image regions into features, and a language model finally outputs text. With only 0.9B parameters, the model achieves the current state-of-the-art OCR accuracy of 94.5% on the standard document-understanding benchmark OmniDocBench v1.5 [21]. The principal motivation for including Paddle1.5 is that it represents the current mainstream visual-language-model OCR architecture and contrasts in design paradigm with traditional rule-based engines, making it a necessary component for completeness of the candidate set. PP-OCRv5 (hereafter PP5) shares the PaddleOCR framework with Paddle1.5 but uses a traditional four-stage workflow architecture; placing the two

side by side renders the architectural paradigm a single controllable comparison variable. ABBYY FineReader 15 (hereafter ABBYY) is included as a general-purpose commercial baseline. While its image pre-processing pipeline is mature, its OCR model is optimized primarily for Latin-script text, providing a reference point against which the Chinese-specialized engines can be compared.

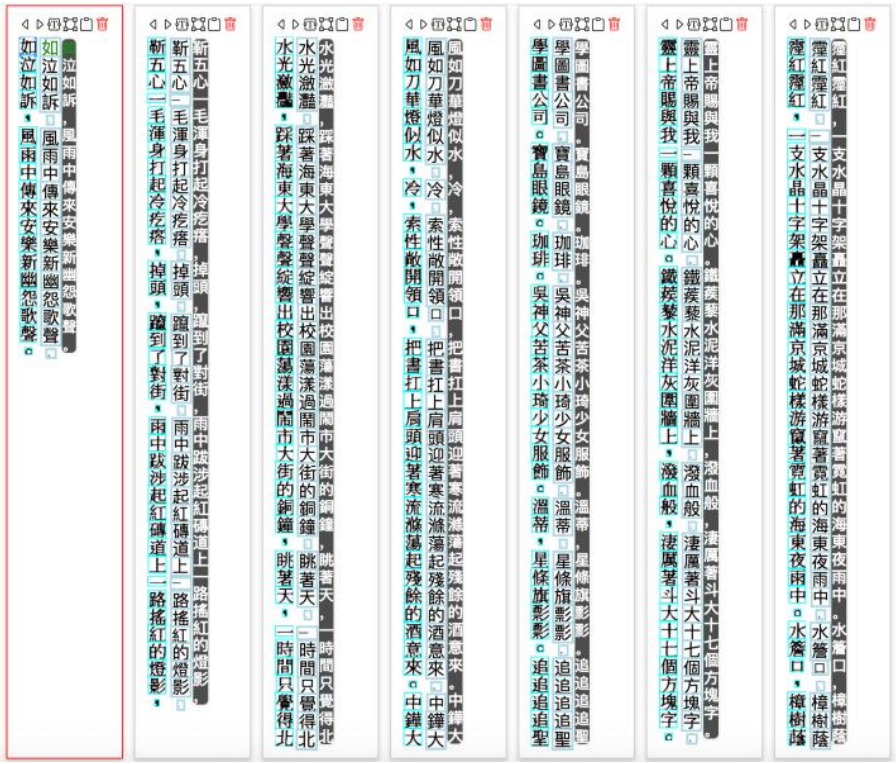

Figure 1. Kandian's intelligent proofreading platform: Character-level three-layer stacked display.
Note: The figure displays several groups of Traditional Chinese text in a three-layer stacked presentation, where the textual content within the different layers of each group is identical.

3.2. Experiment 1: Vertical Traditional Chinese Texts

3.2.1. Materials and Methods

This experiment evaluates three Traditional Chinese literary works: Taiwanese writer-translator Li Yongping's original novel *Haidong Qing: Taibei de Yi Ze Yuyan* (海東青：台北的一則寓言; hereafter *Haidong Qing*) [22], and his Chinese translations of James Redfield's *The Tenth Insight: Holding the Vision* (靈界大覺悟：掌握直覺的新精神世界觀; hereafter *The Tenth Insight*) and Elisabeth Kübler-Ross's *The Wheel of Life: A Memoir of Living and Dying* (天使走過人間：生與死的回憶錄; hereafter *The Wheel of Life*) [23,24], totaling 1,657 PDF pages. All three works are Traditional Chinese, use vertical layout, and contain frequent rare characters. They were published between 1992 and 1998. The original scans show image degradation: yellowed paper, reduced contrast, and blurred edges. These texts represent high-difficulty OCR scenarios in character-set coverage, layout parsing, and image quality. The engines evaluated in this experiment are Kandian, Paddle1.5, and ABBYY; PP5, optimized for standard horizontal printed scenarios and unsuitable for vertical layouts, is not included.

The sampling design follows a purposive equidistant principle, restricted to body-text pages of each work. Sampling does not exclude image-degraded pages, in order to faithfully reflect scan-quality effects on engine accuracy. A baseline of three pages per work was set to ensure minimum coverage, totaling nine pages; the remaining pages were allocated proportionally to each work's page count, subject to a minimum increment of one page per work—*The Tenth Insight* and *The Wheel of Life* each gained one additional page, and the remaining two pages went to the longest work, *Haidong Qing*, which accounts for 57% of total pages. The final sample comprises 13 pages (0.78%), which is comparable to the sampling scales commonly adopted in large-scale historical document digitization projects and provides sufficient grounds for engine-selection decisions [20].

For each sampled page, each engine's OCR text is compared character-by-character against the GT text using edit distance. CER is computed using the editdistance library implementing standard Levenshtein edit distance, with per-page and overall CERs calculated as CER = edit distance (Levenshtein)/|GT| × 100%. Error-type classification is based on the operation sequence extracted by difflib.SequenceMatcher, with each operation assigned to one of three categories: Class A OCR errors—substitutions of non-punctuation characters; Class B OCR errors—substitutions of punctuation characters; and Class C OCR errors—insertions or deletions of any character. The total error count is the sum of the three classes. This classification is applied to two comparison types: character-level differences between each engine's OCR text and the GT text, yielding each engine's total error count; and character-level differences between each engine's OCR text, yielding Class A/B/C divergence count. Full-/half-width punctuation normalization is applied uniformly prior to comparison.

The sample sizes follow controlled-experiment principles. They validate the design logic and the engine-selection decisions, not whole-book accuracy, which requires large-scale sampling after corpus completion. Sample size bounds the precision of any single CER value but not the engine ranking. The cross-engine CER gaps are order-of-magnitude (Kandian 0.25% against Paddle1.5 4.77% and ABBYY 5.96% in Experiment 1), far exceeding the sampling variation a small sample could introduce, so the ranking is stable under resampling.

The following example, drawn from sampled page 2 of *Haidong Qing*, illustrates how the three classes of OCR error map onto operation sequences.

GT: 長長一條**衕衕**黑天裡滿**簷**滿窟兜睞著紅**譻譻**碧熒熒的霓虹。

ABBYY: 長長一條**衛荷**黑天裡滿簷滿窟兜睞著紅**彎膏**碧熒熒的霓虹。

Paddle1.5: 長長一條衕衕黑天裡滿**鶯**滿窟兜睞著紅**蠻蠻**碧熒熒的霓虹。

ABBYY misrecognizes 衕衕 as 衛荷 and 譻譻 as 彎膏, totaling four Class A OCR errors; Paddle1.5 misrecognizes 簷 as 鶯 and 譻譻 as 蠻蠻, totaling three Class A OCR errors. The difflib operation sequence returns replace for all the above substitutions, classified as Class A; neither engine produces Class B or Class C OCR errors on this line.

3.2.2. Results

Per-page CERs and error-type breakdowns for the three engines on the 13 sampled pages, totaling 5,959 characters, are presented in **Tables 1** and **2**.

Table 1. Experiment 1: OCR accuracy validation results (three engines).

Work	ID	Page	Total Chars	Kandian OCR Errors	Kandian CER%	Paddle1.5 CER%	ABBYY CER%
Haidong Qing	H1	188	599	0	0.00	3.84	4.84
Haidong Qing	H2	376	417	1[1]	0.24	5.76	10.07
Haidong Qing	H3	564	466	1[1]	0.21	5.58	5.36
Haidong Qing	H4	752	515	1[1]	0.19	7.57	9.90
Haidong Qing	H5	940	406	0	0.00	7.14	8.37
The Tenth Insight	L1	90	380	0	0.00	7.63	5.79
The Tenth Insight	L2	180	446	1[2]	0.22	2.69	3.59
The Tenth Insight	L3	270	365	0	0.00	6.58	4.93
The Tenth Insight	L4	360	462	0	0.00	2.60	3.03
The Wheel of Life	T1	88	622	6[3]	0.96	2.57	4.82
The Wheel of Life	T2	176	59	1[2]	1.69	6.78	6.78
The Wheel of Life	T3	264	609	1[1]	0.16	4.43	5.25
The Wheel of Life	T4	352	613	3[1]	0.49	3.10	6.20
Overall	—	—	5,959	15	0.25[4]	4.77[5]	5.96[6]

Note: [1] All are Class B OCR errors. [2] All are Class A OCR errors. [3] It includes 2 Class A and 4 Class B OCR errors. [4–6] Overall CER is computed as total edit distance divided by total characters across all sampled pages, not as the arithmetic mean of per-page CERs.

Table 2. Error-type breakdown for the three engines.

Engine	Class A Error	Class B Error	Class C Error	Total Errors	CER%
Kandian	4	11	0	15	0.25
Paddle1.5	95	123	66	284	4.77
ABBYY	105	212	38	355	5.96

The results reveal clear accuracy stratification among the three engines on Traditional Chinese literary text. First, Kandian's overall CER prior to proofreading was only 0.25%, within the 1%–2% reference range on printed

text [5]; after character-by-character human verification, its output served as the GT for this experiment, ensuring sufficient reliability. Second, Paddle1.5 achieved an overall CER of 4.77%, lower than ABBYY's 5.96% but still roughly 19 times Kandian's 0.25%. Third, ABBYY produced substantially more Class B OCR errors at 212, compared to 123 for Paddle1.5, a ratio of 1.72 to 1.

3.3. Experiment 2: Horizontal Simplified Chinese Texts

3.3.1. Materials and Methods

The validation targets of this experiment are the texts in the seven Simplified Chinese sub-corpora, encompassing the Chinese essays of writer-translator Fang Bolin and his Chinese translations of English literary works, Simplified Chinese editions of the originals and translations of Taiwanese writer-translator Li Yongping, and contemporary Simplified Chinese literary works. All texts are Simplified Chinese standard-format texts and differ from Experiment 1's Traditional Chinese literary texts along three dimensions of layout, character set, and image quality, totaling 9,065 pages. Compared with the texts of Experiment 1, the texts in this experiment present significantly lower layout complexity and character-set difficulty; consistency across difficulty levels between the two experiments' conclusions provides additional validation of engine-selection robustness. Given the large total page count, the sampling design follows stratified random sampling: sub-corpora containing multiple works randomly sample one page per work; sub-corpora containing only one work sample two pages. Random numbers were generated using Python's secrets module, yielding 21 pages in total at a sampling rate of approximately 0.23%. The detailed sampling scheme is presented in **Table 3**.

Table 3. OCR accuracy validation sampling plan for seven Simplified Chinese Sub-Corpora.

Sub-Corpus	Work	Total Page	Sampled Page	Rate
Fang's translation of *A Bend in the River*	*A Bend in the River*	295	2	0.68%
Li's translation of *A Bend in the River*	*A Bend in the River*	364	2	0.55%
Fang's translations (extended)	*That Old Ace in the Hole; The Last Testament of Oscar Wilde; All Souls' Day; Let the Great World Spin*	1,762	4	0.23%
Li's translations (extended)	*Mutant Message Down Under; The Celestine Prophecy; The Solitaire Mystery*	1,529	3	0.20%
Fang's originals	*Ah, America!; The Art of Being Mediocre; Essays by an English Learner; A Translator's Rant*	2,162	4	0.19%
Li's originals	*Retribution: The Jiling Chronicles; The End of the River: Mountain*	1,131	2	0.18%
Chinese originals for comparison	*Red Sorghum: A Novel of China; The Right Bank of the Ergun River; Island of Silence; Ancient Capital*	1,822	4	0.22%
Total	—	9,065	21	0.23%

For each sampled page, each engine's OCR text is compared character-by-character against the GT text using the same method as in Experiment 1. The GT is established by Kandian OCR followed by character-by-character human proofreading; the initial CER is 1.01%, higher than Experiment 1's 0.25%, reflecting Kandian's mild sensitivity to Simplified Chinese punctuation variants such as ellipsis and dash forms but not affecting the reliability of the post-proofreading GT.

3.3.2. Results

The error-type breakdown for the four engines on the 21 sampled pages, 12,177 characters total, is presented in **Table 4**.

Table 4. Experiment 2: OCR accuracy validation results (four engines).

Engine	Class A Error	Class B Error	Class C Error	CER%
Kandian[1]	26	90	7	1.01
PP5	13	35	83	1.08
ABBYY	8	109	83	1.64
Paddle1.5	3	13	74	0.74

Note: [1] The row of Kandian reports Kandian's total OCR error count, illustrating the initial OCR accuracy of the GT base text.

The results reveal clear accuracy stratification among the four engines on Simplified Chinese texts. Kandian's

CER prior to human proofreading was 1.01%, and its corrected output serves as the GT for this experiment. Among the three evaluation engines, Paddle1.5 achieved the best accuracy, within Holley's 1%–2% reference range for printed text [5]; PP5 ranked third, also reaching a usable level of accuracy; ABBYY's Class B error count was markedly higher than those of the other engines. Notably, 69 of Paddle1.5's 74 Class C errors are attributable to layout-structural misrecognitions such as page-number boundary overflow, and its character-level OCR accuracy is further improved once such misrecognitions are excluded. Paddle1.5's officially reported document-level overall parsing accuracy of 94.5% on OmniDocBench v1.5 encompasses complex elements including tables and formulas, and is not directly comparable to the character-level CER metric system used in this experiment [21]; the present results constitute an independent measurement on standard Simplified Chinese literary texts, further confirming Paddle1.5's high accuracy in this setting.

3.4. Engine Selection

The two experiments yield consistent patterns across text types, layouts, and difficulty levels, providing a robust empirical basis for engine selection.

In terms of accuracy, Kandian ranked first in Experiment 1 with a pre-proofreading CER of 0.25%, far surpassing Paddle1.5's 4.77% and ABBYY's 5.96%, confirming its specialization for Traditional Chinese OCR. In Experiment 2, with four engines compared, Paddle1.5 achieved the best accuracy at 0.74% CER, followed in declining order by Kandian, PP5, and ABBYY. Across experiments, Paddle1.5 performs consistently in both text-type scenarios.

In terms of limitations, ABBYY exhibits systematically elevated Class B OCR error counts in both experiments—reproduced across Traditional and Simplified Chinese and across high- and low-difficulty levels. PP5 shows the dual limitation of elevated Class A and Class C OCR errors in the Simplified Chinese scenario, and is consistently inferior to Paddle1.5.

In terms of error-pattern, Kandian and Paddle1.5 exhibit inverse distributions across all three OCR error classes: Kandian's 26 Class A OCR errors are 8.7 times Paddle1.5's, and its 90 Class B OCR errors are 6.9 times Paddle1.5's; conversely, Paddle1.5's 74 Class C OCR errors are 10.6 times Kandian's. This complementary distribution ensures that the diff-detection signals between the two engines cover all three error classes, supporting the validity of the multi-engine design.

Based on the above empirical results, this study finalizes the following engine-selection scheme: For Simplified Chinese texts, a multi-engine OCR scheme, combining Kandian and Paddle1.5, is adopted; for Traditional Chinese texts, the wide accuracy gap between engines eliminates the signal validity of diff detection and therefore are not suitable for a multi-engine scheme. Both ABBYY and PP5 are excluded from the final processing workflow. To confirm the exclusion is not a formatting artifact, half-width punctuation normalization and composite dash and ellipsis unification were applied to all four engines, and the outputs were re-scored. The audit records punctuation-level edits only, with no character alterations. ABBYY and PP5 remained non-competitive, since their residual errors are genuine misrecognitions, not removable formatting.

4. Workflow Design

Building on the empirical conclusions of the two experiments, this study establishes a four-stage Simplified Chinese digitization workflow centered on multi-engine collaboration and human-machine complementarity. For the Traditional Chinese scenario, given the different engine-performance profile, this study separately designs a corrective processing path based on Kandian single-engine OCR combined with rule-based post-processing.

The necessity of multi-engine collaboration is empirically supported by Experiment 2. Kandian and Paddle1.5 differ in architectural origin and OCR blind spots, and the non-overlap of their error patterns means that positions where OCR results disagree have a high probability of containing an OCR error, thereby converting human intervention from full-text reading to point-by-point proofreading.

The rationale for human-machine complementarity lies in the inherent limitation of diff detection: when multiple engines produce the same misrecognition for the same character, the error cannot be detected at the comparison stage. To address this, the present study uses two architecturally distinct LLMs—GLM-5 (hereafter GLM) and Claude Sonnet 4.6 (hereafter Claude)—to provide independent verdicts on each disagreement in the divergence list; final human proofreading is then limited to character positions where LLM arbitration is inconsistent and cor-

rectness cannot be automatically determined. Each stage handles only the residual problems left by the previous stage. This avoids repeated full-text scanning.

The workflow consists of four sequential stages, illustrated in **Figure 2**.

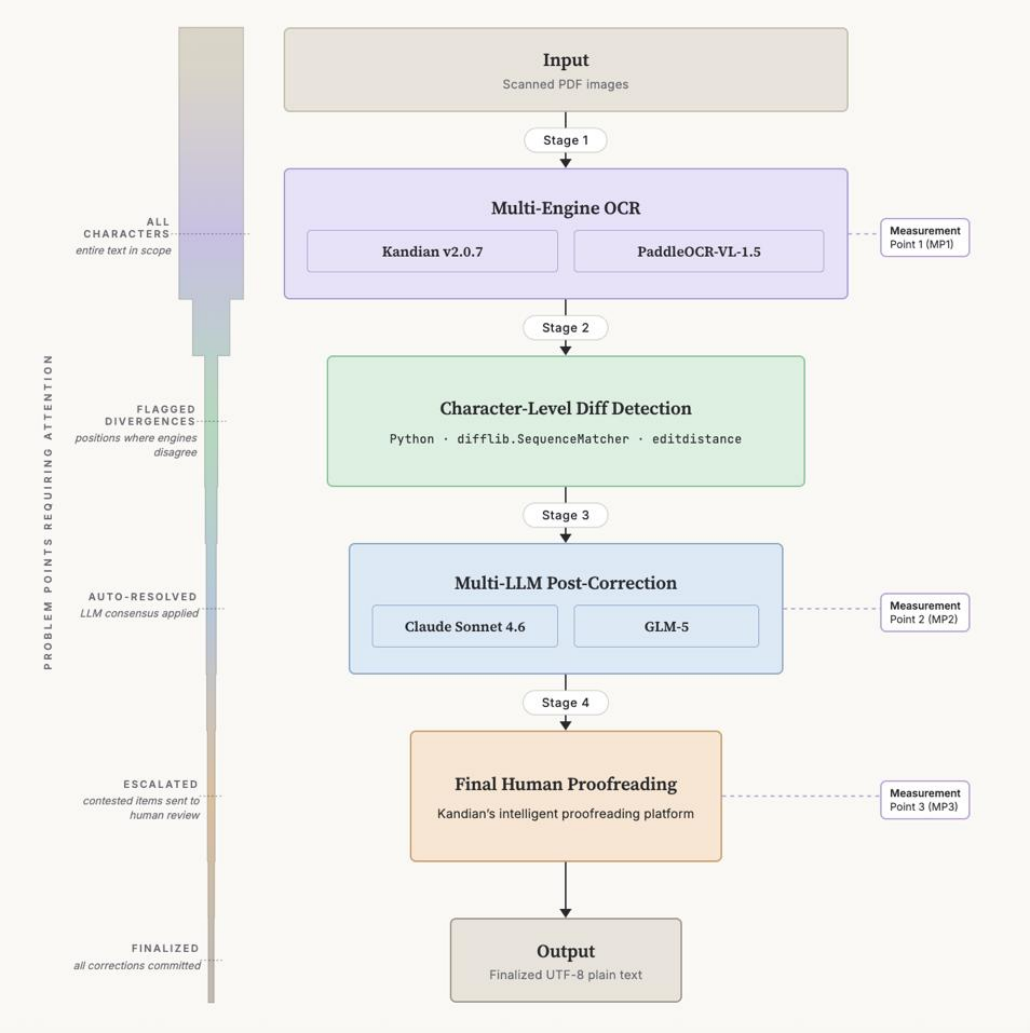

Figure 2. Four-stage digitization workflow for Simplified Chinese literary texts.

Stage 3 operates exclusively on the divergence list produced by Stage 2, so positions where both engines yield the same misrecognition fall outside its reach. Stage 4 therefore performs two functions: it adjudicates divergences on which LLM verdicts conflict or are undeterminable, and it serves as the only safeguard against shared cross-engine misrecognitions that diff detection cannot surface. This stepwise design narrows human involvement from full-text reading to targeted review, keeping proofreading affordable at corpus scale.

5. End-to-End Workflow Validation

5.1. Validation Framework

This section reports workflow-validation evidence: it measures the end-to-end error reduction of the adopted workflow and supports its effectiveness, separate from the engine-comparison evidence in Section 3. This experiment adopts a dual sub-experiment framework to validate the four-stage workflow above. Experiment 3-A targets the horizontal Simplified Chinese text *Hong Gaoliang Jiazu* (红高粱家族; *Red Sorghum: A Novel of China*) from the comparison corpus and executes the complete four-stage workflow [25], examining the workflow's error-reduction in the Simplified Chinese scenario. Experiment 3-B targets the vertical Traditional Chinese text *Haidong Qing* and

executes the Traditional Chinese corrective processing path, serving as a comparative reference to demonstrate the overall workflow's coverage of the Traditional Chinese scenario and to verify whether this path can attain acceptable accuracy without full-text human reading.

Each sub-experiment compares its measurement-point outputs character-by-character against the same independently constructed GT. Experiment 3-A reports three measurement points (post–Stage 1, post–Stage 3, post–Stage 4); Experiment 3-B reports two (post–Kandian OCR, post–rule-based script). The CER changes between adjacent measurement points reflect the contribution of each processing stage. Note that the 20-page sample for Experiment 3-B—Page 3 to Page 22—does not overlap with the sampling for Experiment 1 (H2–H5 corresponding to Page 376/564/752/940), and the two datasets are reported separately.

5.2. Experiment 3-A: Simplified Scenario

5.2.1. Materials and Methods

The validation sample comprises pages 1–20 of the body text of *Hong Gaoliang Jiazu*, with GT established by Kandian OCR followed by page-by-page human proofreading against the original book, totaling 10,464 characters and constructed independently of the workflow.

In Stage 1, Kandian and Paddle1.5 each perform OCR on the same batch of scanned images, producing two independent OCR texts. In Stage 2, Python difflib.SequenceMatcher performs a character-level comparison of the two outputs, grouping one or more consecutive divergent characters into a single divergence as the basic unit for subsequent LLM arbitration; divergences are classified into three types by difflib operation type—Class A, Class B, and Class C OCR errors—and the output is a structured CSV comparison list. In Stage 3, taking the CSV list and the Paddle1.5 OCR text as input, GLM and Claude independently issue verdicts on all divergences; the two sets of verdicts are then merged into three categories: consensus (both models agree—the verdict is applied and the Paddle1.5 text revised accordingly), inconsistent (verdicts conflict), and undeterminable (at least one model declines to decide). Only consensus items are applied at this stage to produce a corrected Paddle1.5 OCR text. The remaining inconsistent and undeterminable items proceed to Stage 4, where each is adjudicated case by case with reference to Kandian's character-level image-text alignment interface and the original scanned images, producing the final OCR text.

CER is measured at three points: the two OCR outputs from Stage 1, the Paddle1.5 OCR text corrected in Stage 3, and the final human-proofread text from Stage 4; each is compared independently against the GT.

5.2.2. Results

Table 5 summarizes the CER changes and error breakdowns at each stage of the workflow.

Table 5. CER changes across the four-stage workflow.

Stage	Class A Error	Class B Error	Class C Error	Total Error	CER%
MP1 Output after Kandian OCR	11	40	5	56	0.54
MP1 Output after Paddle1.5 OCR	14	0	15	29	0.28
MP2 Output after multi-LLM arbitration	12	0	15	27	0.26
MP3 Output after human proofreading	0	0	0	0	0.00

In Stage 2, character-level diff detection identified 62 divergences across the two OCR outputs, transforming downstream proofreading from full-text reading into targeted item-by-item handling. In Stage 3, multi-LLM independent arbitration on the 62 divergences achieved consensus on 49—79.0%. At MP1, the two engines yielded CERs of 0.54% (Kandian) and 0.28% (Paddle1.5). Applying the 49 LLM-consensus corrections to the Paddle1.5 base text reduced CER to 0.26% at MP2. The remaining 13 items—8 with opposed verdicts and 5 undeterminable—proceeded to Stage 4. A human proofreader adjudicated each case using Kandian's character-level image-text alignment interface and independently identified two diff-detection blind spots—local garbled text on Page 17 and a sentence-order misalignment on Page 18, both exhibiting highly consistent misrecognition across engines. These corrections brought MP3 CER to 0.00%. As this result is based on a 20-page sample, actual accuracy at whole-book scale will depend on text characteristics and sampling coverage.

The verdict-type distribution for the 62 divergences is presented in **Table 6** and results for the 13 human-reviewed divergences is presented in **Table 7**.

Table 6. Multi-LLM arbitration result summary.

Verdict	Class A Error	Class B Error	Class C Error	Total
Multi-LLM consensus, Paddle1.5 adopted	16	3	28	47
Multi-LLM consensus, Kandian adopted	2	0	0	2
Multi-LLM disagreement, referred to human review	6	1	1	8
One LLM undeterminable, referred to human review	3	1	1	5
Total	27	5	30	62

Table 7. Multi-LLM arbitration details for the 13 human-reviewed divergences.

ID	Page	Error Class	Claude Verdict	GLM Verdict	Reason	Human Proofreading
6	p.3	A	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
8	p.4	A	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
20	p.7	A	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
25	p.8	B	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
33	p.11	A	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
34	p.12	C	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
44	p.13	A	Kandian correct	Paddle1.5 correct	LLM disagreement	GLM-5 correct
50	p.13	A	Kandian correct	Paddle1.5 correct	LLM disagreement	Claude correct
28	p.9	A	Kandian correct	Undeterminable	GLM uncertain	Claude correct
51	p.14	A	Kandian correct	Undeterminable	GLM uncertain	Claude correct
57	p.17	A	Paddle1.5 correct	Undeterminable	GLM uncertain	Escalated (no consensus)
60	p.18	B	Undeterminable	Paddle1.5 correct	Claude uncertain	Escalated (no consensus)
61	p.18	C	Undeterminable	Paddle1.5 correct	Claude uncertain	Escalated (no consensus)

Among the 8 disagreement cases, Claude's verdict aligned with the human judgment in 7 and GLM's in 1; among the 5 undeterminable cases, Claude's partial verdict was confirmed in 2, while the remaining 3—ID 57, 60, and 61 on pages 17–18—were escalated because neither LLM could supply the correct answer: in each case one model erred and the other declined as undeterminable, so no consensus formed. These three cases, together with two further problems identified during human review—garbled text on page 17 and a sentence-order error on page 18—expose two limits of the automated workflow. First, diff detection flags the divergence, but neither LLM can resolve it; this is a coverage limit, not a consensus error, and the item is correctly routed to human review. Second, the underlying problem was a paragraph-level structural anomaly (garbling, reordered sentences) that diff detection could only surface as scattered character-level divergences, leaving the LLMs unable to recognize the true nature of the error within a character-by-character arbitration frame. Both required the human proofreader to step outside the divergence list. The proofreader returned to the source scans and worked above the character level. Stage 4 is therefore not a residual cleanup step. It handles what falls outside automation's candidate set and granularity.

5.3. Experiment 3-B: Traditional Scenario

5.3.1. Materials and Methods

This experiment uses the first 20 pages, Pages 3–22, of the body text of *Haidong Qing* as the validation sample. The GT, comprising 10,826 characters, was established through Kandian with human proofreading prior to workflow execution. This 20-page sample does not overlap with the *Haidong Qing* pages sampled in Experiment 1.

For the Traditional Chinese scenario, the wide accuracy gap between OCR engines renders the end-to-end workflow inapplicable; instead, a rule-based scripting path is adopted. CER evaluation retains two measurement points: MP1 is the output after Kandian' OCR, and MP2 is the output after rule-based script processing.

The script is implemented in Python under a unified design principle of "character co-occurrence legitimacy": any character pattern that cannot legitimately occur in valid Chinese prose is treated as an OCR error. The script covers three rule categories: half-width punctuation normalization—ASCII commas, periods, question marks, semicolons, colons, and spaces appearing between CJK characters are uniformly replaced with their full-width counterparts or deleted; composite-symbol normalization—various dash and ellipsis variants are unified; encoding normalization—vertical presentation-form punctuation (﹁, ﹂, ﹃ and ﹄, etc.) in the CJK Compatibility Forms block (U+FE30–FE4F) is mapped to its standard horizontal form (「, 」, 『 and 』, etc.). The script additionally

includes several opt-in rules (half/full-width quote mapping, book-title-mark restoration, half/full-width digit normalization, etc.) to accommodate other text types.

All rules share three properties: they rely on character co-occurrence legitimacy rather than semantic judgment, their matching conditions do not depend on the error list of any specific text, and built-in safeguards exempt legitimate uses (decimal points, English words, HTML tags, etc.); the script can therefore be directly extended to other Traditional Chinese vertical literary texts recognized by Kandian. The script generates an audit record for each edit, comprising line number, position, rule, original value, and replacement value.

5.3.2. Results

Table 8 reports the per-page MP1 error distribution across the 20 sampled pages.

Table 8. Per-page MP1 error distribution for Experiment 3-B.

ID	Page	Total Chars	Class A Error	Class B Error	Class C Error	Total Error	CER%
H01	p.3	488	0	0	0	0	0.000
H02	p.4	723	0	0	1	1	0.138
H03	p.5	602	0	0	0	0	0.000
H04	p.6	523	0	0	0	0	0.000
H05	p.7	449	0	0	0	0	0.000
H06	p.8	534	0	1	0	1	0.187
H07	p.9	722	0	0	0	0	0.000
H08	p.10	523	0	1	0	1	0.191
H09	p.11	577	0	3	3	6	1.040
H10	p.12	410	0	0	0	0	0.000
H11	p.13	520	0	2	2	4	0.769
H12	p.14	389	0	0	0	0	0.000
H13	p.15	471	1	0	1	2	0.425
H14	p.16	588	0	0	0	0	0.000
H15	p.17	400	0	0	0	0	0.000
H16	p.18	630	1	0	0	1	0.159
H17	p.19	616	1	1	1	3	0.487
H18	p.20	425	1	1	2	4	0.941
H19	p.21	646	2	3	3	8	1.238
H20	p.22	590	0	0	0	0	0.000
Overall	—	10,826	6	12	13	31	0.286[1]

Note: [1] Overall CER is computed as total edit distance divided by total characters across all 20 sampled pages, not as the arithmetic mean of per-page CERs.

Table 9 summarizes the CER changes and error distributions across the two measurement points of Experiment 3-B.

Table 9. Experiment 3-B: CER summary across two measurement points.

Stage	Class A Error	Class B Error	Class C Error	Total Error	CER%
MP1 Output after Kandian OCR	6	12	13	31	0.286
MP2 Output after rule-based script	6	2	0	8	0.074

The MP1 CER is 0.286%. Class B errors concentrate on dash-formatting problems (the OCR misrecognizes vertical em-dashes as the ASCII pipe |), and Class C errors stem from dash-related character-length mismatches together with three superfluous English periods. After the rule-based script processes the MP1's output, dash normalization matches and replaces 10 instances of "|" with "——", eliminating 10 Class B substitution edits and 10 Class C insertion edits; half-width punctuation normalization removes 3 superfluous ASCII periods, eliminating 3 Class C deletion edits; encoding normalization maps 201 vertical presentation-form quotation marks. The audit log shows that all 13 automatic edits correspond exactly to genuine OCR error positions verified against the GT, yielding 100% precision with no false edits. MP2 CER drops to 0.074%; among the 8 residual errors, 6 are Class A rare-character substitutions and 2 are Class B exclamation-mark formatting issues.

In sum, the rule-based script automatically eliminates 74.2% of the Levenshtein edits in the Traditional Chinese OCR output at 100% precision, reducing CER from 0.286% (MP1) to 0.074% (MP2). The residual errors concentrate on Class A rare-character substitutions, which require image reference for adjudication, and Class B exclamation-mark formatting issues, which require semantic understanding—both lying beyond the limits of rules.

6. Discussion

6.1. Engine Selection

Experiments 1 and 2 yielded consistent regularities under different text types, layouts, and OCR difficulties. ABBYY's systematic misrecognition of the Chinese full-width punctuation system was reproduced in both scenarios; Paddle1.5 outperformed ABBYY in CER in both Traditional (4.77% vs. 5.96%) and Simplified (0.74% vs. 1.64%) Chinese scenarios. PP5 produced markedly more Class C errors than Paddle1.5 in the Simplified Chinese scenario and lacks adaptation to vertical Traditional Chinese, offering no independent selection value overall. This cross-scenario consistency suggests that the selection conclusions generalize robustly to comparable Chinese literary texts. Regarding the prerequisites for multi-engine effectiveness, Experiment 2 confirmed that Kandian and Paddle1.5 exhibit inverse distributions across the three OCR error classes, consistent with the theoretical expectation by Zavorin et al. that multi-engine complementary coverage outperforms single-engine recognition [14]; Experiment 1, by contrast, showed that this prerequisite does not hold for the Traditional Chinese scenario—Kandian and Paddle1.5 differ by nearly 19-fold in CER, eliminating the signal validity of divergence detection.

6.2. Workflow Effectiveness

Experiment 3-A provides a per-stage quantitative basis for the LLM arbitration stage. The architecturally heterogeneous GLM and Claude reached a 79.0% (49/62) verdict consensus on the 62 divergences, indicating that combinations of architecturally distinct LLMs achieve substantial consensus on adjudication tasks; writing the consensus verdicts into the base text reduced MP2 CER from 0.28% to 0.26%, a 0.02-percentage-point reduction, while MP3, after final human proofreading, achieved a further 0.26-percentage-point reduction, exposing a marked asymmetry between the two stages' accuracy contributions. The asymmetry arises because the 62 divergences are dominated by Kandian's Class B OCR errors, while the Paddle1.5 base text had no errors at the punctuation level, leaving little room for direct correction by LLM arbitration. The core contribution of the LLM arbitration stage is therefore routing rather than direct error correction: of 62 divergences, 49 are processed automatically and 13 are escalated to human review, compressing human intervention from full-text reading into targeted adjudication; the 13 LLM-uncertain items were fully corrected after final human proofreading, validating the effectiveness of the human stage as a safety-net mechanism. This is structurally similar to the post-correction framework proposed by Greif et al. [7], with the present study additionally providing per-stage quantitative accuracy data and validating the workflow on Chinese literary text rather than Latin-script historical documents.

The arbitration stage is best read as workload routing rather than direct correction. The 49 consensus verdicts were audited against the GT, and all 49 were correct: no consensus correction was wrong, so the false-consensus count is zero. Arbitration thus routes 79.0% of flagged positions away from human review while introducing no erroneous automatic edit. Relative to a full reading of the 10,464-character sample, human attention is compressed from whole-text proofreading to 62 divergence contexts, of which only 13 require human adjudication. No timing was logged for the arbitration session, so per-item time is not reported; the divergence count itself serves as the workload proxy.

Experiment 3-B presents a processing path and accuracy profile structurally different from the Simplified Chinese scenario: a Python rule-based script grounded in character co-occurrence legitimacy automatically eliminated 74.2% of Levenshtein edits at 100% precision, reducing CER from 0.286% (MP1) to 0.074% (MP2). The function this path performs corresponds structurally to the intermediate filtering function performed by LLM arbitration in the Simplified Chinese scenario, but with a full mechanism that does not involve semantic judgment. The decision to introduce the rule-based path rather than reuse the four-stage workflow rests on two reasons. First, Experiment 1 confirmed that Kandian and Paddle1.5 differ by nearly 19-fold in CER in the Traditional Chinese scenario, invalidating the "complementary error patterns between two engines" prerequisite that divergence detection depends on. Second, in vertical Traditional Chinese literary texts, no reliable statistical boundary at the glyph level distinguishes author-intentional rare characters from OCR misrecognitions; the LLM's reliance on linguistic priors in this scenario carries systematic misjudgment risk—a risk direction consistent with the finding by Kanerva et al. regarding the instability of cross-language correction with a single LLM [19]. Together these point to the conclusion that the effectiveness of LLM post-correction depends heavily on the distinguishability between "correctable characters" and "authorial-style characters" in the target text. The rule-based script trades off coverage for precision

and auditability by limiting automation to error types expressible as character co-occurrence rules (punctuation normalization, encoding normalization); the trade-off is the inability to handle Class A rare-character substitutions and semantically dependent Class B formatting anomalies. The latter constitute the explicit capability boundary of the Traditional Chinese path, manifested in this experiment as the 6 Class A and 2 Class B residual errors at MP2.

6.3. Limitations and Future Directions

This study has four limitations. First, the 20-page sample per sub-experiment in Experiment 3 supports validation of workflow design logic only; whole-book accuracy must be estimated through large-scale sampling after corpus completion. Second, the effectiveness of the LLM combination depends on the versions available at the time of implementation; this paper records all LLM names and versions to support traceability. Third, each processing path has a capability ceiling beyond which automation cannot reach: in the Simplified Chinese scenario, "common blind spots" are an inherent limitation of multi-engine divergence detection; in the Traditional Chinese scenario, Class A rare-character substitutions require image reference and Class B formatting anomalies require semantic judgment, both lying beyond the coverage of rules. The scale of either residual error class can only be estimated through sample-based human proofreading. Fourth, the Traditional Chinese evaluation draws on three works from three Taipei publishers but a single author-translator. The results therefore reflect one set of typesetting and rare-character conventions, and generalizability across authors, periods, and publishers remains untested.

These limitations point to two directions worthy of further pursuit. First, structured feature engineering of the divergence list can be explored by adding contextual n-grams, layout positions, and character types to the existing 62 divergences, in order to test whether the multi-LLM consensus rate can be raised and the final-proofreading workload reduced. Second, cross-genre robustness validation and rule-script extension can be undertaken by applying the Simplified Chinese workflow to academic texts and historical archives, and by extending the Traditional Chinese rule-based script to genres with more complex layouts such as vertical poetry and drama, in order to further delineate the framework's applicability boundaries.

7. Conclusions

This paper constructs and validates a multi-engine OCR evaluation and multi-LLM post-correction framework for Chinese literary corpus construction, and through three controlled experiments fills gaps in the existing literature along three dimensions: the vertical Traditional Chinese modern literary scenario, multi-engine output comparison design, and end-to-end per-stage accuracy quantification.

The engine-evaluation experiments confirm that Kandian holds a substantial accuracy advantage in the vertical Traditional Chinese scenario at 0.25% CER; in the Simplified Chinese scenario, Paddle1.5 leads overall at 0.74% CER, with Kandian (1.01% CER) showing inverse distribution across all three OCR error classes—supporting the multi-engine scheme—while ABBYY is excluded due to systematically elevated Class B errors. The workflow-validation experiment demonstrates that the four-stage Simplified Chinese workflow reduces CER from 0.54% (Kandian) and 0.28% (Paddle1.5) at MP1 to 0.26% at MP2 and 0.00% at MP3, with LLM arbitration functioning as a routing mechanism rather than a direct corrector. In the Traditional Chinese scenario, the wide inter-engine accuracy gap and the indistinguishability between author-intentional rare characters and OCR misrecognitions render the multi-engine workflow inapplicable; a post-processing path based on character co-occurrence legitimacy is adopted instead, reducing CER from 0.286% to 0.074% at 100% rule precision.

These findings show that the layered quality-control approach is operationally feasible and reproducible for Chinese literary text digitization. It uses inter-engine divergence to screen errors and focuses LLM and human proofreading on disagreement positions. The approach can also extend to other Chinese literary genres and to writing systems with large character sets and special layouts. Future research may pursue two directions. One is structured feature engineering of the divergence list and the other is cross-genre robustness validation.

Author Contributions

Conceptualization, X.Z. and R.S.K.C.; Methodology, X.Z.; Software, X.Z.; Validation, X.Z. and R.S.K.C.; Formal Analysis, X.Z.; Investigation, X.Z.; Resources, X.Z.; Data Curation, X.Z.; Writing—Original Draft Preparation, X.Z.; Writing—Review and Editing, R.S.K.C. and R.-H.T.; Visualization, X.Z.; Supervision, R.S.K.C. and R.-H.T.; Project Ad-

ministration, X.Z. and R.S.K.C. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable, as this study analyzed publicly available bibliographic data and did not involve humans or animals.

Informed Consent Statement

Not applicable.

Data Availability Statement

All analysis scripts, the verbatim LLM arbitration prompt template, representative de-identified divergence samples, and the rule-based edit audit logs are openly archived on Zenodo (https://zenodo.org/records/21010627) under CC BY 4.0. The scripts include the OCR CER calculators, the difflib/editdistance divergence-detection script, the rule-based post-processing script, and the four-engine normalization-audit script. Computations use Python 3.9.6 with editdistance 0.8.1 and the standard-library difflib. OCR was performed with Kandian Guji v2.0.7, Paddle1.5, PP5, and ABBYY FineReader 15. Multi-LLM arbitration used Claude Sonnet 4.6 (Anthropic) and GLM-5 (Zhipu AI). Image pre-processing parameters were not logged and are therefore not included.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

For the use of AI Tools, Claude Sonnet 4.6 and GLM-5 were used as post-correction engines within the OCR workflow described in this study, as documented in Sections 4 and 5. The authors used no generative AI tools for manuscript writing. All text was written and verified by the authors, who take full responsibility for the content.

References

1. Eren, F.; Çay, K. Elements of a Digital Urbanisation Strategy for Türkiye: Evidence from Psychometric Testing. *Soc. Sci.* **2025**, *14*, 89.
2. Li, C.; Guo, R.; Zhou, J.; et al. PP-StructureV2: A Stronger Document Analysis System. *arXiv preprint* **2022**, *arXiv:2210.05391*. [CrossRef]
3. Standardization Administration of the People's Republic of China. *GB 18030-2022: Information Technology—Chinese Coded Character Set*; China Standard Press: Beijing, China, 2022.
4. The Unicode Consortium. *The Unicode Standard, Version 17.0.0*; Unicode, Inc.: Mountain View, CA, USA, 2025.
5. Holley, R. How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Mag.* **2009**, *15*. [CrossRef]
6. Neudecker, C.; Baierer, K.; Gerber, M.; et al. A Survey of OCR Evaluation Tools and Metrics. In Proceedings of the 6th International Workshop on Historical Document Imaging and Processing, Lausanne, Switzerland, 5–6 September 2021; pp. 13–18. [CrossRef]
7. Greif, G.; Griesshaber, N.; Greif, R. Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. *arXiv preprint* **2025**, *arXiv:2504.00414*. [CrossRef]
8. Wang, K.; Yang, F.; Jiang, S. Review of Optical Character Recognition. *Appl. Res. Comput.* **2020**, *37*, 22–24. (in Chinese)
9. Chen, X.; Jin, L.; Zhu, Y.; et al. Text Recognition in the Wild: A Survey. *ACM Comput. Surv.* **2021**, *54*, 42. [CrossRef]
10. Li, M.; Lv, T.; Chen, J.; et al. TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Models. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 13094–13102. [CrossRef]

11. Cao, N.A. Resources and Tools for Corpus Compilation of Translated Literary Texts in Late Qing and Republican Period. *J. Transl. Stud.* **2018**, *2*, 153–168.
12. Liu, C.; Jian, C.; Huang, J.; et al. A Robust and Efficient Algorithm for Chinese Historical Document Analysis and Recognition. *Natl. Sci. Rev.* **2023**, *10*, nwad115. [CrossRef]
13. Wong, N.Y.H. Modernism's Vector: Dataset for a Chinese-Language Journal in Late 1950s Malaya. *J. Open Humanit. Data* **2026**, *12*, 40. [CrossRef]
14. Zavorin, I.; Borovikov, E.; Borovikov, A.; et al. A Multi-Evidence, Multi-Engine OCR System. In Proceedings of the Document Recognition and Retrieval XIV, San Jose, CA, USA, 30 January–1 February 2007. [CrossRef]
15. Nguyen, T.T.H.; Jatowt, A.; Coustaty, M.; et al. Survey of Post-OCR Processing Approaches. *ACM Comput. Surv.* **2022**, *54*, 1–37. [CrossRef]
16. Ramirez-Orta, J.A.; Xamena, E.; Maguitman, A.; et al. Post-OCR Document Correction with Large Ensembles of Character Sequence-to-Sequence Models. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 11192–11199. [CrossRef]
17. Rijhwani, S.; Rosenblum, D.; Anastasopoulos, A.; et al. Lexically Aware Semi-Supervised Learning for OCR Post-Correction. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1285–1302. [CrossRef]
18. Thomas, A.; Gaizauskas, R.; Lu, H. Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages, Torino, Italy, May 2024; pp. 116–121.
19. Kanerva, J.; Ledins, C.; Käpyaho, S.; et al. OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches. *arXiv preprint* **2025**, *arXiv:2502.01205*. [CrossRef]
20. Kettunen, K.; Koistinen, M.; Kervinen, J. Ground Truth OCR Sample Data of Finnish Historical Newspapers and Journals in Data Improvement Validation of a Re-OCRing Process. *LIBER Q.* **2020**, *30*, 1–20. [CrossRef]
21. Cui, C.; Sun, T.; Liang, S.; et al. PaddleOCR-VL-1.5: Towards a Multi-Task 0.9B VLM for Robust In-the-Wild Document Parsing. *arXiv preprint* **2026**, *arXiv:2601.21957*. [CrossRef]
22. Li, Y. *Sea-East Green: A Parable of Taipei*; Unitas Publishing: Taipei, Taiwan, 1992. (in Chinese)
23. Redfield, J. *The Tenth Insight: Holding the Vision*; Li, Y., Trans.; Yuan-Liou Publishing: Taipei, Taiwan, 1997. (in Chinese)
24. Kübler-Ross, E. *The Wheel of Life: A Memoir of Living and Dying*; Li, Y., Trans.; Commonwealth Publishing: Taipei, Taiwan, 1998. (in Chinese)
25. Mo, Y. *Red Sorghum: A Novel of China*; PLA Literature and Arts Publishing House: Beijing, China, 1987. (in Chinese)