

Article

Reframing Historical Text Extraction: A Cross-Pathway Validation of OCR, LLM-Assisted Correction, and Direct Multimodal Transcription

Cláudia M. Viana 

Centre of Geographical Studies, Institute of Geography and Spatial Planning, University of Lisbon,
1600-276 Lisboa, Portugal; claudiaviana@edu.ulisboa.pt; Tel.: +351-21-044-3000

Abstract

Historical document collections are increasingly available as digitised images and PDFs, but their conversion into reliable text remains affected by optical character recognition (OCR) errors, degraded pages, heterogeneous layouts, and domain-specific terminology. This study proposes a pathway-level framework for documenting and comparing conventional OCR, OCR followed by large language model (LLM)-assisted correction, and direct multimodal transcription. The framework is demonstrated using the Portuguese Agricultural and Forestry Surveys (1950–1958). A stratified validation sample of 45 pages was selected by visual quality, page type, and geographic coverage. Outputs were evaluated against manually verified reference transcriptions using content-normalised character error rate (CER) and word error rate (WER), document-condition analysis, paired statistical tests, and an entity-level semantic preservation assessment focused on place names, agricultural terms, and measurement expressions. Under the evaluated model and interface conditions, both LLM-based pathways produced lower mean CER and WER than the conventional OCR baseline, with the lowest values observed for direct multimodal transcription. Semantic preservation was also higher for the LLM-based pathways, although measurement expressions remained the most persistent risk, particularly in table-based pages. Downstream tasks were not directly evaluated. The findings support the framework as a method for validating text-extraction pathways before reuse, rather than establishing a universal ranking of tools.

Keywords: large language models; optical character recognition; multimodal transcription; historical document processing; text extraction; semantic preservation; information extraction



Academic Editors: Kelvin K. L. Wong,
Adrian David Cheok and Dhanjoo
N. Ghista

Received: 1 July 2026
Revised: 20 July 2026
Accepted: 23 July 2026
Published: 25 July 2026

Copyright: © 2026 by the author.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Large-scale digitisation has increased access to historical document collections, but access to scanned images is not equivalent to access to reliable textual data. Incomplete or noisy textual layers limit search, indexing, analysis, and reuse, making the evaluation of textual reliability an essential part of historical document processing. Optical Character Recognition (OCR) has long been the main route from image to text. However, historical documents continue to present difficulties for OCR systems. Paper degradation, low contrast, skewed images, typography, historical orthography, abbreviations, tables, marginal notes, and domain-specific vocabulary all affect recognition accuracy [1–4]. OCR errors also affect downstream tasks. Earlier work in information retrieval and digital libraries has

shown that OCR quality can alter search, ranking, and access to digitised collections [5–8]. More recent work shows that OCR noise affects named entity recognition, entity linking, and other information extraction tasks [9,10]. This means that OCR quality cannot be evaluated solely as a transcription problem. It also affects the reliability of the information that can be extracted from the text. Aggregate transcription accuracy can therefore be misleading if the remaining errors affect entities, numerical values, units, or other information units that structure downstream reuse.

Post-OCR correction has therefore become an important part of historical document processing. Earlier approaches used dictionaries, statistical methods, neural sequence models, and task-specific correction pipelines [2,11]. Recent work has extended this discussion to large language models (LLMs), which can use context to correct OCR errors and improve readability. However, LLM-assisted correction is not a neutral operation. It may alter wording, regularise historical spelling, infer missing text, or change entities and numerical values. Recent evaluations of foundation models for historical transcript correction also suggest that LLM-based correction must be tested empirically rather than assumed to improve all outputs in the same way [12]. Multimodal LLMs add a further possibility. Instead of applying OCR first and using an LLM only for correction, a multimodal model can transcribe directly from the digitised image. This creates a new methodological question: how does direct multimodal transcription compare with workflows that retain a conventional OCR stage? Recent work on multimodal models for historical document transcription and information extraction suggests that this route is becoming increasingly relevant, but the answer cannot be assumed in advance [13]. It depends on document quality, layout, language, terminology, model access, model version, costs, and the kind of reuse expected from the textual output.

This paper addresses this problem by reframing historical text extraction as the evaluation of alternative text-extraction pathways rather than as the selection of a single tool. The aim is to present and demonstrate a documented, reusable validation protocol for comparing OCR-based, LLM-assisted, and multimodal text-extraction pathways under documented processing conditions. The paper makes three contributions. First, it proposes a comparative validation framework that treats conventional OCR, LLM-assisted correction, and direct multimodal transcription as alternative pathways for producing textual layers from historical document images. The same manually verified reference transcriptions are used to evaluate each output. Second, it combines aggregate accuracy metrics with document-condition analysis, using Character Error Rate (CER), Word Error Rate (WER), visual quality class, and page type. Third, it complements CER and WER with a semantic preservation assessment focused on three categories relevant to downstream reuse: place names, agricultural terms, and measurement expressions. The framework is demonstrated using a stratified validation sample drawn from the Portuguese Agricultural and Forestry Surveys produced between 1950 and 1958. The corpus contains 272 municipal surveys and 43,822 pages. It combines narrative descriptions, technical agricultural terminology, tables, numerical values, place names, crop names, and variable visual quality.

Textual layers from digitised collections support search, named entity recognition, retrieval-augmented generation, and geographic information extraction. Because OCR errors affect retrieval and information access [5,8], while extraction and correction processes may alter entities and other relevant information units [9,10,14], pathway validation should assess both aggregate accuracy and semantic reliability.

2. Background: Historical Document Text Extraction and LLM-Based Correction

2.1. Historical Document Text Extraction and OCR Uncertainty

OCR converts document images into machine-readable text. In historical collections, this process is affected by image-capture conditions, binarisation, page curvature, degraded ink, stains, unusual typography, line and column segmentation, and the presence of tables or lists [1,15]. Historical spelling and domain-specific terminology introduce additional uncertainty because OCR systems and their language components are often optimised for contemporary language resources. For information processing, aggregate OCR accuracy alone is insufficient, since recognition errors can distort document ranking, retrieval, access, and subsequent text analysis. Earlier studies documented their effects on ranking and relevance feedback [5], downstream text-processing tasks [6], retrieval in the absence of clean-text ground truth [7], and information access in digital libraries [8]. More recent research has identified related problems in named entity recognition and linking [9,10], as well as in retrieval-augmented generation systems [14]. Taken together, these findings show that OCR output should be treated as an intermediate information object whose quality must be evaluated before reuse.

2.2. Post-OCR Correction and LLM-Assisted Correction

Post-OCR correction attempts to improve text after OCR processing. The field includes rule-based correction, dictionary-based approaches, statistical models, neural sequence-to-sequence methods, and more recent language-model-based correction [2,16]. Mei et al. [16] showed that OCR post-processing can be treated as a specific error-correction task rather than as general text cleaning. Lyu et al. [11] showed that neural post hoc correction can reduce OCR errors in historical corpora while controlling for the similarity between the corrected output and the OCR input. Together, these approaches establish post-OCR correction as a controlled transformation of the source output.

LLMs extend this setting by using broader textual context to address broken words, missing accents, incorrect punctuation, and line-break artefacts. However, contextual correction can also regularise historical language, replace uncommon terms with more frequent alternatives, infer unsupported content when the input is ambiguous, or alter entities and numerical values. In historical and domain-specific collections, improved readability may therefore come at the cost of evidential fidelity. Boros et al. [12] evaluated multiple foundation language models across historical post-correction benchmarks, demonstrating the need to assess their performance across languages, periods, document types, and levels of transcription quality. Accordingly, LLM-assisted correction should be evaluated against reference transcriptions and documented in terms of model name and version, prompt, access mode, processing date, and input/output provenance. Such documentation is particularly important because models, interfaces, and access conditions evolve over time.

2.3. Direct Multimodal LLM Transcription

Multimodal LLMs add a further pathway for historical document processing. Instead of applying OCR first and then correcting the OCR output, a multimodal model can transcribe directly from the digitised image. This changes the workflow from image-to-OCR-to-correction to direct image-to-text transcription. Recent studies suggest that this route is becoming technically relevant. Work on handwritten historical documents has reported strong performance for multimodal LLM transcription and correction when compared with specialised handwritten text recognition (HTR) tools [17]. Greif et al. [13] further examined multimodal LLMs for OCR, OCR post-correction, and named entity recog-

nition in historical documents, showing that these models can support both transcription and information extraction tasks. In tabular historical records, Singh and Middleton [18] showed that layout-aware approaches can improve both text extraction and table reconstruction. These studies indicate that direct multimodal approaches provide an alternative to conventional OCR in some settings.

However, these results do not remove the need for validation. Direct multimodal transcription may produce fluent text whose deviations from the source are difficult to detect when no reference transcription is available. It may also omit uncertain elements, regularise orthography, change numbers and units, or add editorial markers. The relevant question is therefore how the performance of multimodal transcription varies relative to OCR-based pathways across different document conditions. Accordingly, it is evaluated as one pathway among others.

2.4. Evaluation Beyond CER and WER

CER and WER are standard metrics for evaluating OCR and transcription outputs. CER measures the edit distance between the generated text and the reference transcription at character level. WER applies the same logic at word level. These metrics are useful because they are transparent, reproducible, and comparable across pathways. However, CER and WER do not fully capture whether analytically important information has been preserved. In historical collections, some errors are more consequential than others. For instance, an incorrect place name may affect search and geocoding. An incorrect crop name or agricultural term may affect domain-specific information extraction. A changed numerical value, unit, cost, area, or production figure may affect quantitative reuse. These errors may be infrequent but important. They may also occur even when aggregate CER or WER improves. For this reason, this study combines CER and WER with a semantic preservation assessment. The analysis records whether each pathway produced category-relevant semantic issues at the entity level, focusing on three categories relevant to downstream reuse: place names, agricultural terms, and measurement expressions. This evaluation strategy follows the broader argument that OCR and transcription quality should be assessed not only through aggregate similarity to a reference text, but also through the effects of textual errors on retrieval, entity processing, information extraction, and downstream reuse [7,9,10,14]. It is particularly relevant for historical agricultural documents, where the value of the textual layer depends not only on readability, but also on the preservation of place names, agricultural terminology, and measurement expressions.

3. Materials and Methods

The methodology compares three text-extraction pathways generated from the same digitised historical pages (Figure 1). The comparison uses manually verified transcriptions as reference data. Each pathway is evaluated separately: conventional OCR; OCR followed by LLM-assisted correction; and direct multimodal transcription.

3.1. Document Collection

The workflow is demonstrated using the Portuguese Agricultural and Forestry Surveys (Inquéritos Agrícolas e Florestais) produced between 1950 and 1958 in the context of the Portuguese Agricultural Promotion Plan (Figure 2). The surveys were prepared by agronomists and foresters and describe agricultural and forestry conditions at the municipal scale. The available collection comprises 272 municipal surveys and 43,822 pages. Each survey includes narrative text, tables, technical descriptions, and, in many cases, hand-drawn cartographic materials. The internal structure usually includes an agronomic survey, a forestry survey, and a section on municipal problems. The content covers climate,

soils, hydrology, transport infrastructure, crops, crop rotation systems, fertilisers, pest control, agricultural industries, production, consumption, forestry, erosion control, and land suitability. The collection combines historical language, domain-specific terminology, variable visual quality, continuous text, tables, numerical content, and complex layouts, making it suitable for pathway-level validation.

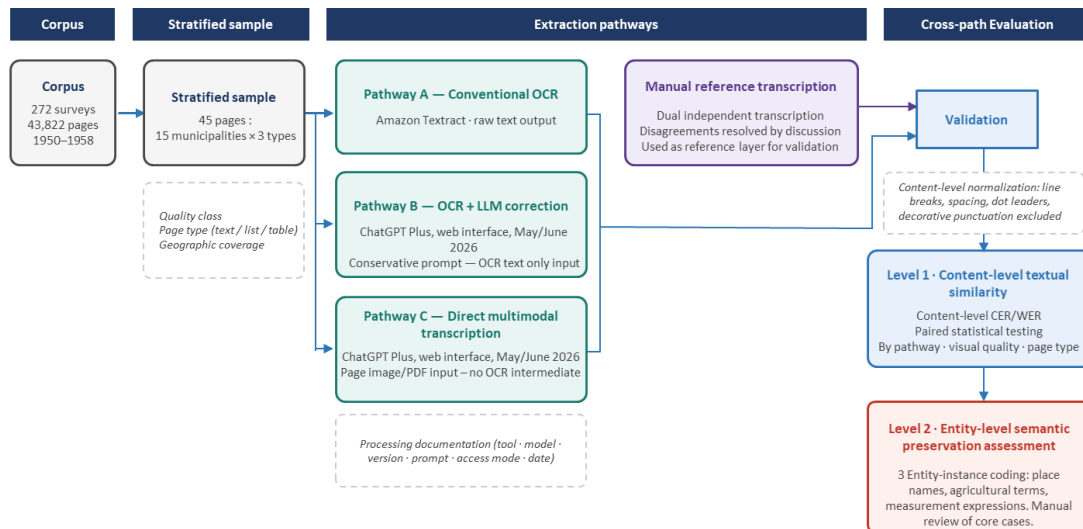


Figure 1. Comparative validation workflow for historical document text extraction.

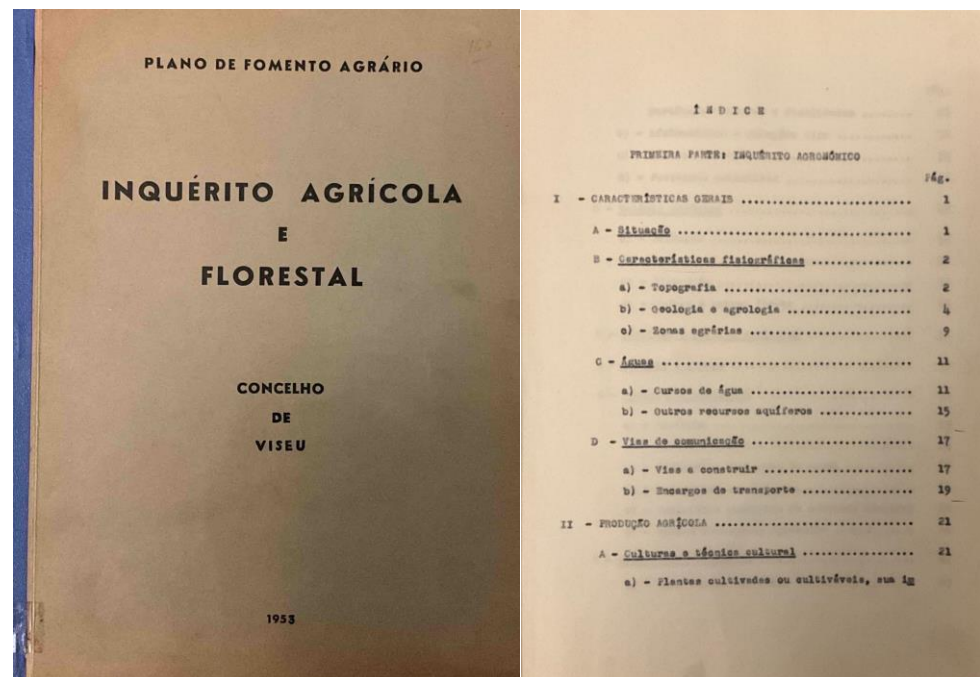


Figure 2. Example pages from the Portuguese Agricultural and Forestry Surveys (Inquéritos Agrícolas e Florestais, 1950–1958), showing the cover and table of contents of the volume for the municipality of Viseu, published in 1953. The original Portuguese text is retained as part of the historical source. The figure is included to illustrate the appearance and documentary structure of the analysed materials.

3.2. Corpus and Validation Sample

Before text extraction, the collection was organised at the municipal survey level. Each survey was treated as a documentary unit and assigned an identifier. The identifier links the physical source, digitised image/PDF, OCR output, LLM-corrected output, direct

multimodal transcription, and validation records. Digitisation was carried out before the current comparative experiment. The original workflow used Adobe Scan to capture pages and apply contrast enhancement and a two-tone black-and-white filter. A stratified validation sample was created to assess pathway performance under different document conditions. The digitised pages had previously been classified into three visual quality classes: high, medium, and low (Figure 3). This classification was conducted manually through qualitative visual inspection of the digitised page images. No automated image-quality metrics, numerical scores, or formulae were used. High-quality pages were those with clear printed text, high contrast between text and background, limited staining or visual noise, and no major obstruction to reading. Medium-quality pages were those with readable text but with moderate contrast loss, uneven illumination, mild blur, small stains, page curvature, or localised degradation. Low-quality pages were those with weak contrast, degraded paper, blurred or broken characters, strong stains, visual noise, difficult line segmentation, or tables and layout features that made text recognition uncertain. The classification was used to ensure variation in the validation sample, not as an automated image-quality metric.

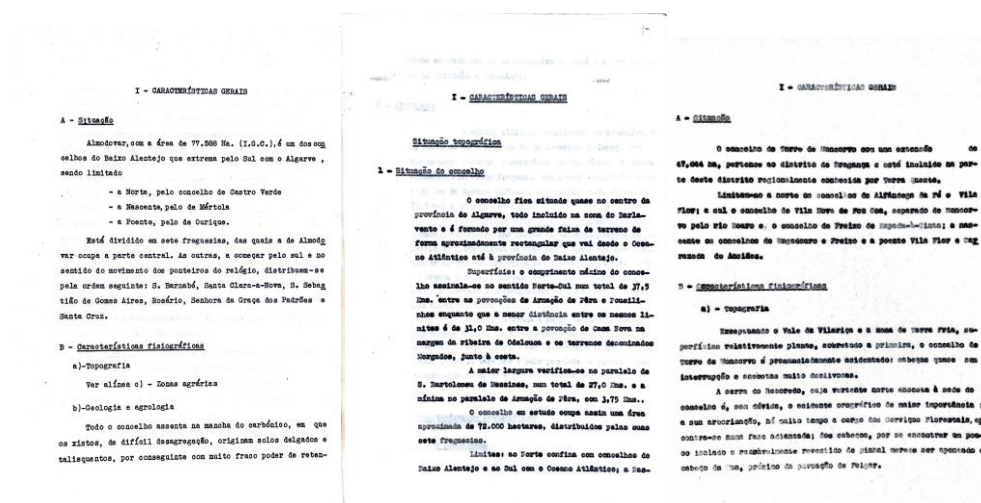


Figure 3. Illustrative examples of high-, medium-, and low-quality pages used in the visual classification of the validation sample. The classification considered characteristics such as text clarity, contrast, background interference, page degradation, and print irregularities. The original Portuguese text is retained because the panels reproduce unmodified pages from the historical documents. Supplementary File S3 provides the 45 digitised page PDFs included in the validation sample.

The 45-page validation sample was constrained by the time and cost required to produce manually verified ground truth. It was designed primarily for pathway comparison rather than corpus-level estimation, prioritising validation depth over corpus-wide coverage. The sampling design was stratified but non-probabilistic. The municipal surveys and pages were selected purposively to capture planned variation in visual quality, page type, and geographic coverage; no random selection was performed within strata. The validation sample comprised 15 municipal surveys, with three pages selected from each survey. At page level, the sample was balanced by visual quality, with 15 high-quality, 15 medium-quality, and 15 low-quality pages. The selected surveys were geographically distributed to avoid concentrating the sample in a small number of municipalities or regions. Within each selected survey, one page was selected for each of three analytical page-type categories. T1 comprised continuous descriptive text. T2 included agroecological or technical terminology related to subjects such as crops, soils, water resources, or production systems. T3 comprised more complex content, including layout problems, tabular

information, numerical values, abbreviations, or low legibility. The municipality, survey and page identifiers, visual quality class, page type, and selection rationale for each of the 45 pages are provided in Table S1 of Supplementary File S2.

3.3. Text-Extraction Pathways

The experiment compares three text-extraction pathways (Table 1). Pathway A uses a pre-existing conventional OCR layer generated with Amazon Textract during an earlier operational phase of the AgroecoDecipher project. The selection of Amazon Textract followed comparative testing on documents from the same historical collection, as reported in [3]. The OCR layer was generated through the AWS Management Console using the DetectDocumentText–OCR feature, with digitised PDFs uploaded from local storage and the resulting TXT files retained as the raw OCR output [3]. In the present study, this layer is treated as a fixed operational baseline rather than as a benchmark of OCR engines. The three pathways were evaluated as complete alternative workflows rather than as components of a single composite model. The paired A–B comparison isolates the contribution of LLM-assisted correction applied to the same OCR output, whereas Pathway C represents a separate image-to-text workflow that bypasses the conventional OCR layer.

Table 1. Input, processing step, and output of the three text-extraction pathways.

Pathway/Layer	Input	Processing Step	Output	Main Question
A. Conventional OCR	Digitised page image/PDF	Amazon Textract OCR	Raw OCR text	What is the baseline performance of a scalable conventional OCR system?
B. OCR + LLM-assisted correction	Same raw OCR text from Pathway A	Conservative LLM-assisted post-OCR correction	LLM-corrected OCR text	Does LLM-assisted correction improve the same OCR input without changing key information?
C. Direct multimodal transcription	Digitised page image/PDF	Direct image-to-text transcription by multimodal LLM	Direct multimodal transcription	How does direct multimodal transcription compare with the OCR-based pathways under the evaluated conditions?

The LLM-assisted correction (Pathway B) was treated as a controlled text transformation rather than as automatic cleaning. The prompt was designed to be conservative. It instructed the model to correct evident OCR errors, preserve historical vocabulary, maintain structure where possible, keep place names, agricultural terms, numbers, units, and measurement expressions unchanged unless there was clear evidence of OCR error, and mark uncertain words rather than infer them. For direct multimodal transcription, the model received the corresponding digitised page image/PDF and did not receive the Amazon Textract OCR output or the manually verified reference transcription. The direct multimodal transcription prompt followed the same conservative principle.

Both the OCR + LLM-assisted correction and direct multimodal transcription outputs were generated using the ChatGPT web application under a ChatGPT Plus subscription (OpenAI), accessed between May and June 2026. GPT-5.5 Thinking was manually selected for all processing sessions and was the software/model version displayed in the model picker. The web interface did not expose backend model identifiers or inference parameters, and therefore could not be controlled or recorded. The prompts and processing documen-

tation template are provided in Supplementary File S1, while the linked page-level inputs, outputs, and reference transcriptions are provided in Supplementary File S4.

3.4. Reference Transcriptions

For each validation page, a manually verified reference transcription was produced by the lead researcher and used as ground truth for calculating error metrics [19,20]. The reference transcription was intended to preserve the original text as far as possible, including historical spelling, punctuation, accentuation, numerical values, units, place names, agricultural terms, and measurement expressions where these could be read with confidence. A second researcher then independently transcribed the same page without access to the lead researcher's transcription or to any of the outputs generated by the three text-extraction pathways. Once both transcriptions had been completed, they were compared page by page. Differences were identified through direct comparison. The two transcriptions were in full agreement for 41 of the 45 pages. The remaining four disagreements involved measurement expressions, specifically numerical values and table-based quantities where the source image was ambiguous or partially legible. In these cases, both researchers returned to the source image and discussed the affected passages. The agreed version was retained as the reference transcription, and the error metrics for the affected pages were recalculated accordingly. Line wrapping and unambiguous visual word fragmentation were normalised, meaning that the reference transcription serves as content-level ground truth rather than as a reconstruction of page layout.

3.5. Content-Based Textual Validation and Statistical Comparison

Two quantitative metrics were used to evaluate the textual outputs: Character Error Rate (CER) and Word Error Rate (WER). Both metrics are based on edit distance, following the Levenshtein formulation of insertions, deletions, and substitutions [21]. CER applies this comparison at the character level, while WER applies it at the word-token level. Both metrics are widely used for evaluating OCR, handwriting recognition, speech recognition, and transcription quality [20,22]. Lower values indicate better performance. CER is calculated as

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c} \quad (1)$$

where S_c , D_c , and I_c are the number of character-level substitutions, deletions, and insertions required to transform the generated output into the reference transcription, and N_c is the total number of characters in the reference transcription.

WER is calculated as

$$\text{WER} = \frac{S_w + D_w + I_w}{N_w} \quad (2)$$

where S_w , D_w , and I_w are the number of word-level substitutions, deletions, and insertions required to transform the generated output into the reference transcription, and N_w is the number of word tokens in the reference transcription.

Before calculating CER and WER, all outputs were converted to plain text and normalised using the same rules. Repeated spaces, tabs, blank lines, line breaks, indentation, decorative separators, and punctuation used only for visual alignment were removed. Accents, alphabetic characters, numerical values, measurement units, and content-relevant punctuation were preserved. WER was calculated on normalised word tokens, with punctuation and whitespace excluded as independent tokens. Because historical documents often contain line-wrapped or visually fragmented words, adjacent split tokens were treated as equivalent to their joined form when their concatenation matched the corresponding token. For example, *con + celhos* was treated as equivalent to “*concelhos*” [municipalities]. Tables and lists were evaluated as linearised text; table reconstruction, column alignment, and

visual spacing were not assessed. Errors affecting numerical values, units, crop names, place names, technical terms, or the interpretation of table values were examined separately in the semantic preservation audit.

For each validation page, CER and WER were calculated separately for Pathway A, Pathway B, and Pathway C. Relative error reduction was then calculated for Pathways B and C relative to the conventional OCR baseline. Results are reported for the full 45-page sample, by visual quality class, and by page type. Because the same 45 reference pages were processed through all three extraction pathways, pathway comparisons were conducted at the page level using paired analyses. For both CER and WER, three planned comparisons were performed: Pathway A versus Pathway B, Pathway A versus Pathway C, and Pathway B versus Pathway C. Paired reductions were calculated as the difference in error rate between the first and second pathway in each comparison, so that positive values indicate lower error in the second pathway.

Given the sample size and the skewed distribution of page-level error rates, Wilcoxon signed-rank tests were used to assess paired differences for the three pathway comparisons (A–B, A–C, and B–C), separately for CER and WER. A positive paired reduction indicates lower error for the second pathway. Mean and median paired reductions were reported with 95% percentile bootstrap confidence intervals calculated from 10,000 resamples of the paired page-level differences. Relative mean reductions and their bootstrap confidence intervals were reported as interpretable effect-size measures. Wilcoxon *p*-values were adjusted jointly across the six planned comparisons using the Holm procedure. The Wilcoxon test evaluates the signed ranks of the paired differences, whereas the bootstrap confidence interval quantifies uncertainty around the arithmetic mean paired reduction. Because these procedures assess different summaries of the paired differences, a statistically significant Wilcoxon result may coexist with a bootstrap confidence interval for the mean that includes zero, particularly when the differences are skewed or heterogeneous. The page-level validation dataset and statistical comparison outputs are provided in Supplementary File S2.

3.6. Entity-Level Semantic Preservation Assessment

The entity-level semantic preservation assessment examined whether information units relevant to downstream reuse were preserved across the three text-extraction pathways. Three semantic categories were audited: place names, agricultural terms, and measurement expressions (Table 2).

Table 2. Semantic categories included in the entity-level preservation assessment.

Semantic Category	Includes	Reason for Inclusion
Place names	Municipalities, parishes, rivers, mountains, valleys, regions, and other named locations	Relevant for search, geocoding, spatial linking, and geographic information extraction
Agricultural terms	Crop names, fruit trees, agricultural products, crop varieties, farming practices, and agrarian or forestry terminology	Relevant for agricultural information extraction and historical interpretation
Measurement expressions	Numerical values with explicit or implicit units, including areas, production values, costs, percentages, intervals, population counts, and table-based quantities	Relevant for quantitative reuse and interpretation of tables, production values, areas, and costs

The assessment followed a semi-automated extraction and review procedure (Table 3). First, candidate entities were extracted from the manually verified reference transcriptions. Place-name candidates were identified using capitalised multi-word expressions and a controlled list of municipalities, parishes, hydrological features, regions, and other named locations observed in the validation pages. Agricultural-term candidates were identified using a controlled vocabulary compiled from the reference transcriptions, including crop names, fruit trees, agricultural products, crop varieties, forestry terms, and domain-specific agrarian expressions. Measurement-expression candidates were extracted using rule-based patterns for numerical values, currency expressions, percentages, ranges, units, and table-based quantities. These patterns captured integers and decimal values, thousands separators, intervals, monetary expressions ending in \$00, and units such as hectares, kg, kg/ha, litres, metres, kilometres, percentages, and population counts.

Table 3. Workflow for the entity-level semantic preservation assessment.

Step	Procedure	Output
Candidate extraction	Entities were extracted from reference transcriptions using controlled vocabularies for place names and agricultural terms, and rule-based patterns for measurement expressions.	Initial candidate entity list.
Pathway alignment	Each reference entity was aligned with the corresponding output in Pathways A, B, and C using content-normalised matching.	Pathway-specific entity comparison.
Discrepancy cross-check	Candidate differences were checked against the final CER/WER discrepancy logs.	Candidate semantic issues.
Output-only detection	Category-relevant entities appearing in outputs but not alignable to the reference were flagged as hallucination candidates.	Output-only candidate list.
Cleaning	Generic headers, layout artefacts, OCR fragments, isolated page markers, and invalid category assignments were excluded.	Cleaned semantic audit dataset.
Manual review	Mandatory, measurement-expression, quality-control, and optional cases were reviewed manually.	Final valid and invalid candidate labels.
Final aggregation	Valid reference entities were used as the denominator; hallucinations were reported separately.	Preservation, alteration, omission, uncertainty, and hallucination summaries.

Each reference candidate was aligned with the corresponding outputs from Pathways A, B, and C using content-normalised string comparison. Differences in spacing, line breaks, decorative punctuation, and unambiguous visual word fragmentation were allowed during alignment. Candidate discrepancies were cross-checked against the final CER/WER discrepancy logs. The same category rules were applied to the pathway outputs to identify output-only entities that could not be aligned with the reference. These cases were treated as hallucination candidates and reviewed manually.

Each valid reference entity instance received one code for each pathway: Preserved, Altered, Omitted, Hallucinated, or Uncertain. Preserved indicated that the information unit retained the same content; Altered indicated a change that could affect interpretation; Omitted indicated absence from the output; Hallucinated indicated an output-only entity; and Uncertain was used when the reference or alignment could not be resolved confidently.

The candidate set was cleaned by excluding generic table headers, page markers, layout artefacts, OCR fragments, and invalid category assignments. Manual review covered all mandatory cases, all measurement-expression discrepancies, and a stratified quality-control sample of automatically preserved or low-risk cases. In total, 372 rows were manually reviewed. Final summaries were calculated from valid reference entity instances, while retained output-only hallucinations were reported separately and excluded from the preservation-rate denominator.

The final semantic audit retained 988 valid reference entity instances after adjudication. To assess coding reliability, a stratified sample of 99 instances, corresponding to approximately 10% of the final audit, was independently coded by a second researcher using the same classification protocol. Raw agreement was 87.9% (87 of 99 instances). Disagreements primarily involved borderline measurement expressions and ambiguous numerical digits in low-legibility tables and were resolved through adjudication.

The complete semantic audit protocol, including extraction rules, coding definitions, exclusion criteria, risk classifications, and coded examples, is provided in Supplementary File S1. The validation data and semantic preservation summaries are provided in Supplementary File S2.

4. Results

4.1. Overall Pathway Performance

Across the 45-page validation sample, the conventional OCR baseline produced a mean CER of 11.74% and a mean WER of 29.89% (Table 4). Both LLM-based pathways reduced the error rates relative to this baseline. Pathway B reduced the mean CER to 4.74% and the mean WER to 8.47%, corresponding to relative reductions of 59.67% and 71.66%, respectively. Pathway C produced mean error rates of 3.84% for CER and 6.68% for WE, corresponding to relative reductions of 67.27% and 77.65%, respectively, compared with Pathway A.

Table 4. Overall CER and WER performance by pathway.

Pathway	Mean CER	Mean WER	Relative CER Reduction vs. A	Relative WER Reduction vs. A
A. Conventional OCR	11.74%	29.89%	-	-
B. OCR + LLM-assisted correction	4.74%	8.47%	59.67%	71.66%
C. Direct multimodal transcription	3.84%	6.68%	67.27%	77.65%

Pathways B and C both had significantly lower rates than the OCR baseline after Holm correction (all adjusted $p < 0.001$) (Table 5). The B–C comparisons also remained statistically significant after Holm correction (CER: adjusted $p = 0.033$; WER: adjusted $p = 0.011$). However, the absolute mean and median paired reductions were small, particularly for CER, and the bootstrap confidence intervals for the mean paired reductions included zero. The B–C difference should therefore be interpreted as statistically detectable but modest and heterogeneous across pages. Full page-level CER and WER results are provided in Supplementary File S2.

4.2. Performance by Visual Quality

Error rates varied substantially across visual quality classes (Figure 4a,b). The OCR baseline had its lowest error rates on high-quality pages and deteriorated as visual quality decreased. The two LLM-based pathways reduced error rates in all three visual quality classes. For high-quality pages, Pathway A produced a mean CER of 3% and a mean WER

of 6%. Pathway B reduced these values to 2% and 3%, while Pathway C produced the lowest mean error rates, with a CER of 1% and a WER of 2%.

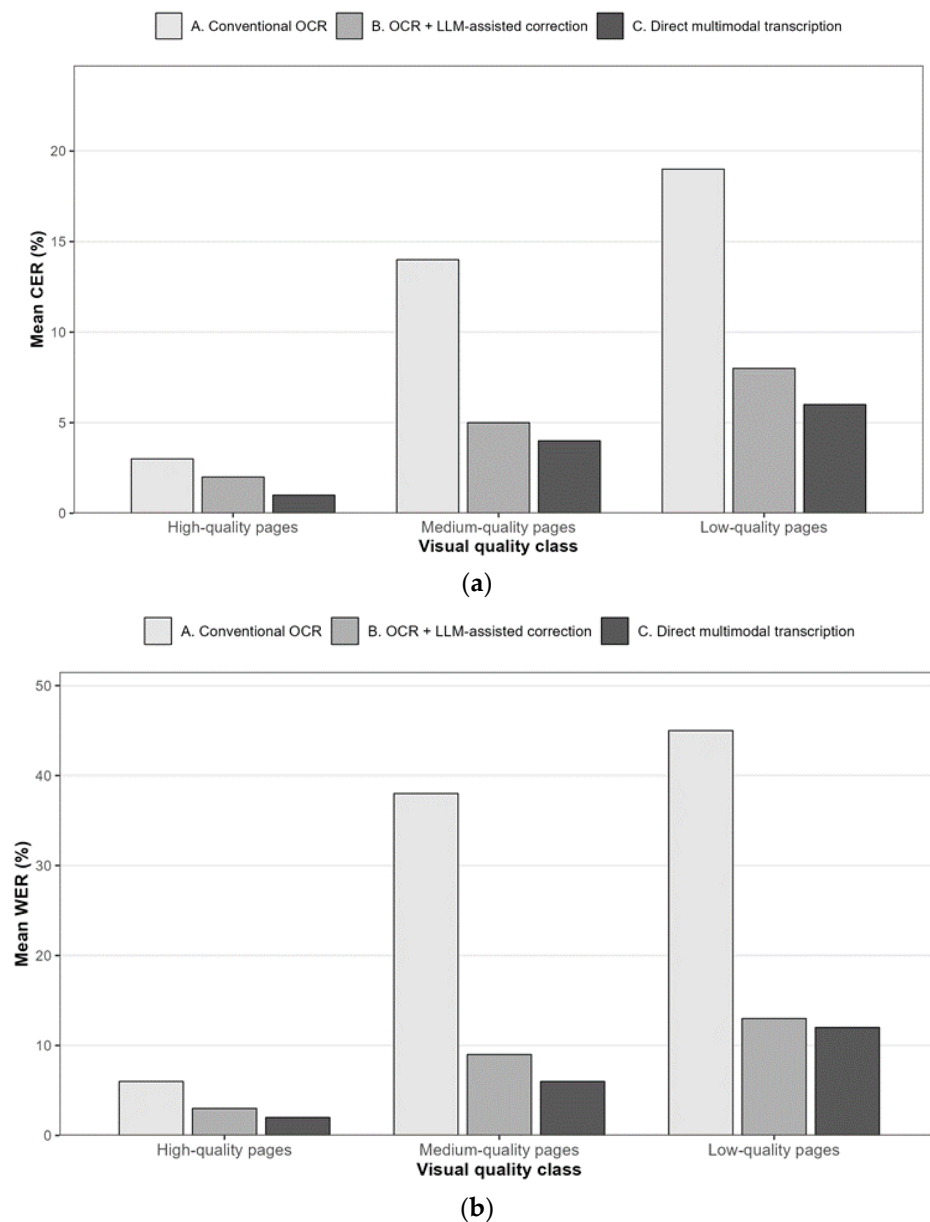


Figure 4. (a) Mean Character Error Rate (CER) by visual quality class and text-extraction pathway; (b) Mean Word Error Rate (WER) by visual quality class and text-extraction pathway.

Table 5. Paired statistical comparison of CER and WER reductions between pathways. Positive reductions favour the second pathway in each comparison. Relative mean reduction is reported as an interpretable effect-size measure. Values outside parentheses in the final column are unadjusted *p*-values; values within parentheses are Holm-adjusted *p*-values calculated across the six planned comparisons.

Metric and Comparison	Mean Paired Reduction, Percentage Points (95% Bootstrap CI)	Median Paired Reduction, Percentage Points (95% CI)	Relative Mean Reduction, % (95% CI)	Wilcoxon <i>p</i> -Value (Holm-Adjusted)
CER: A vs. B	7.01 (5.12–8.98)	5.17 (2.07–8.89)	59.7 (46.4–74.3)	<0.001 (<0.001)
CER: A vs. C	7.90 (5.59–10.30)	5.14 (1.96–9.66)	67.3 (52.6–81.6)	<0.001 (<0.001)

Table 5. *Cont.*

Metric and Comparison	Mean Paired Reduction, Percentage Points (95% Bootstrap CI)	Median Paired Reduction, Percentage Points (95% CI)	Relative Mean Reduction, % (95% CI)	Wilcoxon <i>p</i> -Value (Holm-Adjusted)
CER: B vs. C	0.89 (−1.27–2.94)	0.29 (0.00–0.53)	18.8 (−36.1–51.4)	0.033 (0.033)
WER: A vs. B	21.42 (16.40–26.37)	19.77 (9.70–29.45)	71.7 (63.1–79.3)	<0.001 (<0.001)
WER: A vs. C	23.21 (17.07–29.17)	19.77 (8.54–33.79)	77.6 (65.3–86.7)	<0.001 (<0.001)
WER: B vs. C	1.79 (−1.35–4.41)	1.34 (0.57–2.65)	21.1 (−17.8–47.2)	0.006 (0.011)

For medium-quality pages, the OCR baseline deteriorated substantially, reaching a mean CER of 14% and a mean WER of 38%. Pathway B reduced these values to 5% and 9%. Pathway C produced a mean CER of 4% and a mean WER of 6%. Both LLM-based pathways therefore reduced the effect of increasing page degradation, with Pathway C showing lower mean error rates than Pathway B within this class.

For low-quality pages, Pathway A had the highest error rates, with a mean CER of 19% and a mean WER of 45%. Pathway B reduced these values to 8% and 13%. Pathway C produced a mean CER of 6% and a mean WER of 12%. The largest gains occurred on medium- and low-quality pages, although residual errors in low-quality and table-based content reinforce the need for semantic preservation checks.

4.3. Performance by Page Type

Performance also varied by page type. For continuous descriptive pages (T1), Pathway A produced a mean CER of 8% and a mean WER of 24%. Pathway B reduced these values to 2% and 4%. Pathway C produced similar character-level performance and the lowest word-level error rate, with a mean CER of 2% and a mean WER of 3% (Table 6). Both LLM-based pathways were therefore effective for continuous prose. For pages containing technical or agroecological terminology (T2), Pathway A produced a mean CER of 8% and a mean WER of 29%. Pathway B reduced these values to 1% and 4%. Pathway C produced a mean CER of 1% and a mean WER of 3% in this group. Both LLM-based pathways substantially improved terminology-rich pages, with only a small difference in mean error rates between Pathways B and C. Pages with complex layouts, numerical content, or tables (T3) had the highest error rates across all pathways. In this group, Pathway A produced a mean CER of 19% and a mean WER of 37%. Pathway B reduced these values to 12% and 17%. Pathway C produced mean values of 9% for CER and 14% for WER. Lower aggregate error rates do not necessarily imply safe quantitative reuse, because small changes in digits, units, table positions, or categories may have serious interpretative consequences. These pages therefore require semantic preservation checks in addition to CER and WER.

Table 6. CER and WER performance by page type.

Page Type	A CER	A WER	B CER	B WER	C CER	C WER
Continuous text pages	7.97%	24.19%	1.56%	4.33%	1.58%	2.87%
Technical/agroecological terminology pages	7.84%	28.70%	1.13%	4.12%	0.89%	3.42%
Complex layout/numerical/table pages	19.42%	36.77%	11.51%	16.96%	9.07%	13.74%

4.4. Entity-Level Semantic Preservation Results

The semantic preservation assessment retained 988 valid reference entity instances: 143 place names, 442 agricultural terms, and 403 measurement expressions (Table 7). Seven valid output-only hallucinations were identified for Pathway A, while none were

retained for Pathways B or C. Preservation increased from 73.58% for Pathway A to 95.45% for Pathway B and 96.15% for Pathway C. The difference between the two LLM-based pathways was therefore small, but both showed substantially higher preservation rates than the OCR baseline.

Table 7. Overall entity-level semantic preservation outcomes by pathway.

Pathway	Valid Reference Entities	Preserved	Altered	Omitted	Output-Only Hallucinations	Preservation Rate	High-Risk Errors
A. Conventional OCR	988	727	214	47	7	73.58%	250
B. OCR + LLM-assisted correction	988	943	43	2	0	95.45%	37
C. Direct multimodal LLM transcription	988	950	37	1	0	96.15%	38

The largest improvements occurred for place names and agricultural terms (Table 8). Place-name preservation increased from 72.73% in Pathway A to 92.31% in Pathway B and 97.90% in Pathway C, while agricultural-term preservation increased from 71.49% to 97.74% and 98.42%, respectively. Measurement expressions remained the most sensitive category: Pathway A preserved 76.18%, compared with 94.04% for Pathway B and 93.05% for Pathway C. Residual high-risk errors were concentrated in altered numerical values, missing or altered units, and table-based quantities. Pathway C had the highest preservation rates for place names and agricultural terms, whereas Pathway B had a slightly higher preservation rate for measurement expressions. These results show that lower CER and WER do not necessarily guarantee safe quantitative reuse.

Table 8. Entity-level semantic preservation by semantic category and pathway.

Semantic Category	Pathway	Valid Entities	Preserved	Altered	Omitted	Preservation Rate	High-Risk Errors
Place names	A	143	104	38	1	72.73%	34
Place names	B	143	132	11	0	92.31%	5
Place names	C	143	140	3	0	97.90%	3
Agricultural terms	A	442	316	124	2	71.49%	121
Agricultural terms	B	442	432	9	1	97.74%	8
Agricultural terms	C	442	435	7	0	98.42%	7
Measurement expressions	A	403	307	52	44	76.18%	95
Measurement expressions	B	403	379	23	1	94.04%	24
Measurement expressions	C	403	375	27	1	93.05%	28

The detailed coding of measurement expressions showed that Pathway A combined numerical alterations with frequent omissions, whereas the fewer residual issues in Pathways B and C primarily involved numerical value changes (Table 9). Even isolated changes in digits, ranges, units, or table values may affect quantitative interpretation, supporting the separate validation of numerical and table-based content. Detailed semantic preservation summaries are provided in Supplementary File S2.

Table 9. Main types of semantic issues observed in measurement expressions after manual review.

Issue Type	Description	A	B	C	Downstream Risk
Numeric digit/value change	One or more digits or numerical values were changed.	48	22	27	Distorts quantitative interpretation and numerical extraction.
Omission	A measurement expression present in the reference was omitted from the output.	44	1	1	Reduces recall for quantitative reuse and may remove relevant table values.
Unit change	A unit or measurement marker was changed or not preserved with the value.	3	0	0	Reduces interpretability of areas, production values, costs, or other quantities.
Range change	A numerical interval or range was altered.	1	1	0	Can change the interpretation of thresholds, classes, or grouped values.

5. Discussion

5.1. From OCR Tool Selection to Pathway Comparison

The main contribution of this study is methodological rather than technological. Instead of benchmarking individual OCR engines or model versions, the framework treats historical text extraction as a pathway-level validation problem under explicit processing conditions and reuse objectives. This perspective is particularly relevant for workflows combining commercial OCR services, interface-mediated LLM systems, and multimodal tools whose models and access conditions may change over time. Reproducibility therefore depends on documenting inputs, prompts, processing conditions, output layers, validation procedures, and intended reuse contexts. This pathway-oriented framing is consistent with Greif et al. [13], who describe multimodal LLMs as providing early evidence of a potential paradigm shift in historical data collection and document transcription, while also recognising that no single study can exhaustively evaluate all languages, fonts, layouts, and document conditions. The present study contributes to this emerging perspective by focusing on how alternative text-extraction workflows can be documented and validated, rather than on identifying a universally superior engine or model.

Both LLM-based pathways produced lower aggregate error rates than the OCR baseline, while direct multimodal transcription yielded the lowest mean CER and WER within the validation sample. However, the magnitude of these improvements varied according to page quality and content type. This is consistent with evidence that multimodal LLMs can support historical transcription, OCR correction, and information extraction [13,17,23], although performance varies across languages, document types, model configurations, and prompts [12,24,25]. Prompts, processing dates, input types, output layers, normalisation rules, and validation records should therefore form part of the documented workflow.

Pathway selection should consequently be treated as a context-dependent methodological decision. Collections dominated by continuous text may require different strategies from those containing degraded pages, complex layouts, or numerically dense tables. Factors such as font type, image resolution, information density, degradation, and prompt design may influence transcription accuracy [13,25]. Validation should therefore focus on pathway behaviour under documented conditions rather than on claims of universal superiority.

5.2. Aggregate Accuracy and Semantic Reliability

CER and WER were necessary for comparing pathway performance, but they did not fully represent the reliability of the outputs for reuse. This is consistent with previous work showing that OCR noise affects retrieval, document access, named entity recognition, entity linking, and other downstream information-processing tasks [7–10]. In the present study, lower aggregate error rates generally corresponded to fewer semantic issues, but this relationship was not absolute.

The entity-level assessment showed that both LLM-based pathways substantially improved the preservation of place names and agricultural terms. Measurement expressions nevertheless remained the most sensitive category because small changes in numerical values, units, costs, areas, percentages, production values, or table-based quantities may substantially affect interpretation. The difficulty of this content is consistent with research showing that historical table digitisation and structured numerical extraction require validation beyond aggregate transcription metrics [18,26].

A transcription can therefore be highly readable while remaining unreliable for extracting quantities, costs, crop values, or areas. Similarly, correction may improve CER and WER while changing a place name, agricultural term, or numerical expression. Generative variability and hallucinations remain relevant concerns, and LLM-based correction should be evaluated empirically across languages, periods, document types, and transcription-quality levels [24]. Semantic preservation assessment consequently complements rather than replaces CER and WER.

Digit substitutions, unit changes, omitted quantities, and table-value displacement identify plausible risks for geocoding, information extraction, and quantitative historical analysis. However, these downstream tasks were not directly evaluated. Task-based assessment is required to determine whether, and to what extent, such errors affect downstream outputs.

5.3. Implications for Historical Document Workflows

Historical document workflows should not treat OCR, LLM-assisted correction, and multimodal transcription as mutually exclusive options. Conventional OCR remains useful as a scalable and auditable intermediate layer. Pathway B substantially improved this baseline, particularly at the word level, but occasionally normalised spelling, changed wording, or altered specific information units. It should therefore be treated as a controlled transformation rather than as neutral cleaning. Previous research similarly shows that LLM-based correction does not generalise uniformly across historical datasets and that performance depends on language, document type, model choice, and prompt configuration [12,24].

Within the validation sample, Pathway C produced lower aggregate error rates than Pathways A and B, particularly on medium- and low-quality pages, but residual numerical and table-related errors confirm that multimodal transcription does not remove the need for validation. Previous work has also reported potential time and cost advantages for LLM-based transcription relative to specialised HTR services, although these aspects were not evaluated in the present study [17].

Recent benchmark studies indicate that no consistent overall winner emerges across document types and languages and that performance may be weaker for open-source models and non-English scripts [24,25,27]. Document-processing performance also varies across layouts, scripts, and information structures [28,29]. An adaptive workflow is therefore appropriate: continuous-text pages may be suitable for direct multimodal transcription; terminology-rich pages may be processed through either LLM-based pathway but require checks of domain-specific terms; and table-based or numerically dense pages require ad-

ditional validation regardless of pathway. The optimal approach depends on processing scale, accuracy requirements, traceability, and available resources.

5.4. Limitations and Future Work

The validation design was diagnostic rather than exhaustive. The 45-page sample was purposively stratified by visual quality, page type, and geographic coverage but cannot represent every condition in the 43,822-page collection. The paired design supports internal pathway comparison because all outputs were evaluated on the same pages, but the findings should not be interpreted as population-level performance estimates. The 988 semantic entities constitute a separate entity-level audit and do not increase the effective sample size of the page-level CER/WER tests.

Visual quality was classified through direct inspection rather than automated image-quality indicators. The study also evaluated text extraction and semantic preservation rather than end-to-end downstream tasks such as named entity recognition, retrieval, question answering, geocoding, or retrieval-augmented generation. In addition, the comparison reflects the specific OCR service, LLM interface, displayed model label, prompt, and processing period used. Because the ChatGPT web interface did not expose the backend model build or inference parameters, the reported label documents the access conditions rather than a fully specified computational environment. More generally, the commercial interfaces did not expose their underlying computational infrastructure, limiting the evaluation to observable inputs and outputs.

The semantic audit combined semi-automated candidate extraction with explicit category definitions, rule-based detection of measurement expressions, controlled vocabularies, output-only detection, and human review of high-risk and quality-control cases. These procedures support reproducibility at the level of coding decisions while recognising that historical spelling, OCR noise, and table structure still require expert judgement. The use of the same manually verified references for all pathways supports the internal validity of the comparison.

Process recoverability means that the inputs, transformations, outputs, and review evidence can be reconstructed from the retained records; it does not imply that a commercial services can be rerun under identical conditions. The retained objects, prompts, normalisation rules, coding decisions, and validation records support reanalysis, while the original interface-mediated executions remain only partially reproducible.

Future work should apply the framework to larger and more diverse collections and established public benchmarks, evaluate downstream tasks directly, and compare additional pathway configurations using open-source OCR engines and locally deployable vision–language models, such as Qwen2.5-VL [30]. Locally controlled implementations would allow for transcription accuracy and semantic preservation to be examined alongside computational requirements, including FLOPs, GPU memory use, and scalability with image resolution. These experiments would help distinguish tool- and collection-specific effects from patterns that recur across pathway configurations, while providing more stable computational environments for independent replication.

6. Conclusions

This study demonstrates that, within the evaluated validation sample and processing conditions, both LLM-based pathways produced lower CER and WER than the fixed conventional OCR baseline. Direct multimodal transcription yielded mean CER and WER values of 3.84% and 6.68%, respectively, while OCR followed by LLM-assisted correction yielded corresponding values of 4.74% and 8.47%. Performance nevertheless varied according to visual quality and page type, with complex layouts, numerical content, and tables

remaining challenging across pathways. The semantic preservation assessment further showed that textual accuracy and semantic reliability are related but not equivalent. Place names and agricultural terms were generally well-preserved by the LLM-based pathways, whereas measurement expressions remained particularly sensitive. The results therefore do not establish a universal ranking of text-extraction approaches.

The main contribution of the study is a structured framework for documenting, comparing, and validating text-extraction pathways in historical collections. The framework combines manually verified reference transcriptions, content-normalised CER and WER, paired statistical testing, document-condition analysis, semantic preservation assessment, and processing documentation. It supports an adaptive workflow in which conventional OCR provides a scalable baseline, LLM-assisted correction is treated as a controlled transformation, and direct multimodal transcription remains subject to validation for structured and quantitative content. The findings are limited to the historical collection, OCR service, LLM access mode, model label, prompt, and processing period evaluated. Downstream tasks were not directly tested. Future work should apply the framework to larger and more diverse collections and examine how the resulting textual layers affect task-based outcomes.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/info17080722/s1>. Supplementary File S1 provides the methodological protocol, prompts, processing documentation template, semantic candidate extraction rules, semantic preservation codebook, exclusion and risk-classification rules, and examples of coded cases. Supplementary File S2 provides the validation dataset, including page-level sample metadata, CER and WER results for each pathway, paired statistical comparison outputs, and semantic preservation summaries. Supplementary File S3 contains the 45 digitised page PDFs included in the validation sample. Supplementary File S4 provides the 180 page-level textual objects used in the comparison: 45 conventional OCR outputs (Pathway A), 45 LLM-corrected OCR outputs (Pathway B), 45 direct multimodal transcriptions (Pathway C), and 45 manually verified reference transcriptions (REF). It also includes a file manifest, naming crosswalk, and SHA-256 checksums. Each textual object is linked through Page_ID to the corresponding source-page PDF provided in Supplementary File S3. Table S1: Sample_page_level.

Funding: This work was funded by national funds through FCT—Fundação para a Ciência e a Tecnologia, I.P.—under project 2022.09372.PTDC and through the Centre for Geographical Studies, Universidade de Lisboa, under grant UID/00295/2025 (<https://doi.org/10.54499/UID/00295/2025>).

Data Availability Statement: The data supporting the findings of this study are available in the Supplementary Materials accompanying this article.

Acknowledgments: The author thanks the AgroecoDecipher project team for their collaboration and support throughout the project. Special thanks are due to Jorge Rocha for independently verifying the reference transcriptions and entity-level coding. The author also acknowledges GEOMODLAB—Laboratory for Remote Sensing, Geographical Analysis and Modelling—of the Centre of Geographical Studies/IGOT for providing the equipment and software required for this study.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings. *Appl. Sci.* **2019**, *9*, 4853. [CrossRef]
2. Nguyen, T.T.H.; Jatowt, A.; Coustaty, M.; Doucet, A. Survey of post-OCR processing approaches. *ACM Comput. Surv.* **2021**, *54*, 124. [CrossRef]
3. Viana, C.M.; Carvalho, D. Unveiling historical agroecological patterns through artificial intelligence (AI) and Geographic Information Systems (GIS). *AGILE GISci. Ser.* **2024**, *5*, 49. [CrossRef]

4. Viana, C.M. Reflexão sobre as abordagens geoespaciais da investigação geográfica aplicadas à modelação dos sistemas agrícolas. *Finisterra* **2023**, *58*, 181–196. [CrossRef]
5. Taghva, K.; Borsack, J.; Condit, A. Effects of OCR errors on ranking and feedback using the vector space model. *Inf. Process. Manag.* **1996**, *32*, 317–327. [CrossRef]
6. Lopresti, D. Optical character recognition errors and their effects on natural language processing. *Int. J. Doc. Anal. Recognit.* **2009**, *12*, 141–151. [CrossRef]
7. Ghosh, K.; Chakraborty, A.; Parui, S.K.; Majumder, P. Improving information retrieval performance on OCRred text in the absence of clean text ground truth. *Inf. Process. Manag.* **2016**, *52*, 873–884. [CrossRef]
8. Chiron, G.; Doucet, A.; Coustaty, M.; Visani, M.; Moreux, J.-P. Impact of OCR errors on the use of digital libraries: Towards a better access to information. In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, Toronto, ON, Canada, 19–23 June 2017; pp. 249–252. [CrossRef]
9. Hamdi, A.; Linhares Pontes, E.; Sidere, N.; Coustaty, M.; Doucet, A. In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Nat. Lang. Eng.* **2023**, *29*, 425–448. [CrossRef]
10. Ehrmann, M.; Hamdi, A.; Linhares Pontes, E.; Romanello, M.; Doucet, A. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.* **2024**, *56*, 27. [CrossRef]
11. Lyu, L.; Koutraki, M.; Krickl, M.; Fetahu, B. Neural OCR post-hoc correction of historical corpora. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 479–493. [CrossRef]
12. Boros, E.; Ehrmann, M.; Romanello, M.; Najem-Meyer, S.; Kaplan, F. Post-correction of historical text transcripts with large language models: An exploratory study. In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 133–159. [CrossRef]
13. Greif, G.; Griesshaber, N.; Greif, R. Multimodal LLMs for OCR, OCR post-correction, and named entity recognition in historical documents. *arXiv* **2025**, arXiv:2504.00414.
14. Song, M. Defining the problem: The impact of OCR quality on retrieval-augmented generation performance and strategies for improvement. *Inf. Process. Manag.* **2026**, *63*, 104368. [CrossRef]
15. Boros, E.; Hamdi, A.; Linhares Pontes, E.; Cabrera-Diego, L.A.; Moreno, J.G.; Sidere, N.; Doucet, A. Alleviating digitization errors in named entity recognition for historical documents. In Proceedings of the 24th Conference on Computational Natural Language Learning; Fernández, R., Linzen, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 431–441. [CrossRef]
16. Mei, J.; Islam, A.; Moh'd, A.; Wu, Y.; Milios, E.E. Statistical learning for OCR error correction. *Inf. Process. Manag.* **2018**, *54*, 874–887. [CrossRef]
17. Humphries, M.; Leddy, L.C.; Downton, Q.; Legace, M.; McConnell, J.; Murray, I.; Spence, E. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Hist. Methods* **2025**, *58*, 175–193. [CrossRef]
18. Singh, L.G.; Middleton, S.E. Tabular context-aware optical character recognition and tabular data reconstruction for historical records. *Int. J. Doc. Anal. Recognit.* **2025**, *28*, 357–376. [CrossRef] [PubMed]
19. Springmann, U.; Reul, C.; Dipper, S.; Baiter, J. Ground truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *J. Lang. Technol. Comput. Linguist.* **2018**, *33*, 97–114. [CrossRef]
20. Neudecker, C.; Baierer, K.; Gerber, M.; Clausner, C.; Antonacopoulos, A.; Pletschacher, S. A survey of OCR evaluation tools and metrics. In Proceedings of the 6th International Workshop on Historical Document Imaging and Processing; Association for Computing Machinery: New York, NY, USA, 2021; pp. 13–18. [CrossRef]
21. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
22. Morris, A.C.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In Proceedings of the Interspeech 2004, Jeju Island, Republic of Korea, 4–8 October 2004; pp. 2765–2768. [CrossRef]
23. Narbona, A.; Ros, S. Beyond traditional OCR: Exploring the efficiency of LLMs in document processing. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025); CEUR Workshop Proceedings, Zaragoza, Spain, 23 September 2025. Available online: https://ceur-ws.org/Vol-4098/PastReader2025_paper1.pdf (accessed on 1 May 2026).
24. Kanerva, J.; Ledins, C.; Käpyaho, S.; Ginter, F. OCR error post-correction with LLMs in historical documents: No free lunches. In Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains; University of Tartu Library: Tartu, Estonia, 2025; pp. 38–47. Available online: <https://aclanthology.org/2025.resourceful-1.8/> (accessed on 1 May 2026).
25. Levchenko, M.A. Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities; INCOMA Ltd.: Shoumen, Bulgaria, 2025; pp. 75–85. Available online: <https://aclanthology.org/2025.lm4dh-1.7/> (accessed on 1 May 2026).
26. Bäcker-Peral, V.; Meursault, V.; Severen, C. Can LLMs Credibly Transform the Creation of Panel Data from Diverse Historical Tables? Federal Reserve Bank of Philadelphia Working Paper No. 25-28. *arXiv* **2025**, arXiv:2505.11599.

27. Ouyang, L.; Qu, Y.; Zhou, H.; Zhu, J.; Zhang, R.; Lin, Q.; Wang, B.; Zhao, Z.; Jiang, M.; Zhao, X.; et al. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville TN, USA, 11–15 June 2025; pp. 24838–24848.
28. Kersapati, M.I.; Perono Cacciafoco, F.; Sihite, B.; Wu, S.; Widyaningrum, K.P.; Atqa, M.; Toni, E.A.B. Unravelling lexical and narrative patterns in the Hikayat Lonthoir: A computational linguistics approach. *Information* **2025**, *16*, 1069. [\[CrossRef\]](#)
29. Bray, N.; Hempel, M.; Boeding, M.; Sharif, H. Decoding technical diagrams: A survey of AI methods for image content extraction and understanding. *Information* **2026**, *17*, 165. [\[CrossRef\]](#)
30. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.13923. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.