Beyond Traditional OCR: Exploring the Efficiency of LLMs in Document Processing

Antonio Narbona[1,†], Salvador Ros[2,*,†]

[1]Universidad de Alcalá de Henares, UAH, Spain. Centro de Estudios Universitarios Ramón Areces
[2]Universidad Nacional de Educación a Distancia, UNED, Spain

Abstract

This work explores two distinct OCR architectures for historical document processing, comparing a complex, multi-stage pipeline with a streamlined approach based solely on large language models (LLMs). Our findings demonstrate that the LLM-based architecture significantly outperforms the more traditional, complex systems, achieving superior accuracy with lower Character and Word Error Rates (CER and WER). The results highlight that, without the need for supplementary preprocessing or fusion steps, LLMs represent the current state-of-the-art in OCR, offering both high precision and efficiency. This approach marks a new era in historical document transcription, establishing LLMs as the most effective solution for heritage digitization workflows in the digital humanities.

Keywords

OCR, paper template, Large Language Models, Document Processing

1. Introduction

The digitization and analysis of historical documents written in Spanish from the 17th to the 20th century are of critical importance for historical, linguistic, and cultural research. These primary sources offer unique insights into the evolution of language, social customs, and the historical context of past centuries. Optical Character Recognition (OCR) technology plays a pivotal role in the transformation of these physical documents into machine-readable text, enabling large-scale access, searchability, and computational analysis. Recent advances in artificial intelligence have significantly enhanced the potential of OCR, facilitating the automated processing of fragile archival materials without physical manipulation and converting them into structured, searchable digital formats.

Despite these advancements, applying OCR to historical documents presents persistent challenges, Manrique-Gómez et al. (2024). The physical degradation of paper—such as yellowing, brittleness, and ink fading—combined with printing inconsistencies, often complicates accurate character recognition. Historical Spanish texts frequently employ archaic or decorative typefaces that deviate markedly from modern fonts, and handwritten marginalia or annotations further impede recognition accuracy. These issues commonly result in OCR errors, including character misrecognition, incorrect punctuation, and the introduction of extraneous symbols or artifacts.

The quality of OCR output directly affects the efficiency and accuracy of subsequent text correction workflows. High error rates require more intensive post-processing, which can be time-consuming and computationally costly, particularly when correction systems struggle with distinguishing OCR-induced noise from valid but rare historical language patterns. Moreover, significant linguistic challenges arise when working with pre-modern Spanish. Between the 17th and 19th centuries, the Spanish language underwent considerable orthographic and grammatical evolution. Variations in spelling, archaic vocabulary, obsolete verb forms, and shifting syntactic norms present obstacles for both OCR systems and modern natural language processing (NLP) tools.

IberLEF 2025 September 2025, Zaragoza, Spain.

[*] Corresponding author.

[†] These authors contributed equally.

anarbona@yahoo.com (A. Narbona); sros@scc.uned.es (S. Ros)
0000-0002-9421-8566 (S. Ros)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-10

Addressing these challenges requires correction methodologies that are both linguistically informed and computationally efficient. Large Language Models (LLMs) offer promising capabilities in this domain. Users increasingly seek to leverage LLMs to correct OCR outputs with high precision while adapting to historical language variation. Promising directions include the use of prompt-based correction methods, which are cost-effective and fast to deploy, versus fine-tuning strategies, which may yield greater accuracy at the cost of higher computational demands. The optimal approach will depend on the user's constraints, including processing time, accuracy requirements, and available resources. Recent work in historical OCR correction combines rule-based postprocessing with deep learning approaches, leveraging domain-specific corpora, adaptive tokenization, and few-shot learning paradigms. These hybrid models show improved performance in tasks involving diachronic language variation and high noise levels typical of historical documents Subramani et al. (2021). In this study, we focus on the PastReaders task, conducted within the framework of IberLEF2025, González-Barba et al. (2025), a shared evaluation initiative aimed at benchmarking Natural Language Processing systems for the Spanish language Montejo-Ráez et al. (2025b).

2. Related Work

The Spanish language of the 17th to 19th centuries exhibits a dynamic and evolving character, with significant grammatical and orthographic differences from its modern counterpart. Orthographic variants such as *avía* instead of *había*, *fierro* instead of *hierro*, and *estoi* instead of *estoy* were common. Accentuation rules also diverged, with accents appearing more frequently or being placed differently. While grammatical evolution was less drastic, variations in verb conjugation and pronoun usage—such as the occasional use of *vosotros* in parts of Latin America—highlight the linguistic heterogeneity of the period.

These historical linguistic characteristics pose a considerable challenge for Optical Character Recognition (OCR) systems. The interaction between OCR-induced typographic noise and authentic historical variation complicates efforts to accurately reconstruct the original text. Moreover, systematic orthographic transformations further confound correction efforts. The Historical Ink project has cataloged recurrent graphemic alternations, such as: *á* and *a* (e.g., *hara* → *hará*), *é* and *e* (*fué* → *fue*), *í* and *i* (*decia* → *decía*), *ó* and *o* (*ocasion* → *ocasión*), *ú* and *u* (*ningun* → *ningún*), *i* and *y* (*mui* → *muy*), *j* and *g* (*jente* → *gente*), *v* and *b* (*gravado* → *grabado*), *s* and *x* (*espiró* → *expiró*), *j* and *x* (*méjico* → *méxico*), *c* and *s* (*faces* → *fases*), and *s* and *z* (*dies* → *diez*).

A robust correction system must be intelligent enough to distinguish these historical variants from spurious OCR artifacts. Misidentifying genuine linguistic features as errors may result in overcorrection and the unwanted modernization of texts, compromising their historical authenticity. The methodology developed by the Historical Ink project is particularly useful in this context, as it classifies changes—such as accentuation—as surface-level features rather than errors, thereby preserving typographic norms relevant to the period, Manrique-Gómez et al. (2024).

Early attempts to address this problem relied on hybrid systems, handcrafted rules, and prompt engineering to guide general-purpose models. These methods sought to compensate for the inability of early language models to parse historical variation in Spanish. Although prompt-based strategies could yield satisfactory results when contextual parameters were carefully defined, these systems often failed to distinguish legitimate constructions from OCR noise, leading to distortion or loss of original forms, Boros et al. (2024).

Fine-tuning large language models (LLMs) on corpora like Latam-XIX marked a significant improvement. Models trained on historical data better preserved grammatical and orthographic patterns, but the approach remained resource-intensive and was limited by access to adequate training data and computational infrastructure, Yenigün (2025). Specialized systems developed in the Historical Ink project further improved precision by learning to differentiate between diachronic variation and OCR noise.

Despite these advances, a paradigm shift is underway. Recent state-of-the-art LLMs demonstrate the ability to perform high-quality corrections directly from raw OCR input using only well-formulated prompts,Greif et al. (2025), Kim et al. (2025). These models consistently match or exceed the performance of fine-tuned or hybrid systems without requiring extensive adaptation. Our findings confirm that these next-generation models possess a remarkable capacity to understand typographic noise, recognize historical forms, and preserve textual authenticity—thus rendering many traditional support techniques obsolete. In this context, the role of LLMs transitions from compensating for model limitations to capitalizing on their generalizability and contextual fluency. Their simplicity, efficiency, and precision establish a new standard in the digitization and correction of historical Spanish texts.

3. Methodology

In the present work we use a methodological approach grounded in the need to reconcile the inherent complexity of historical Spanish texts with the evolving capacities of OCR and large language models. At the heart of this process lies a progressive rethinking of architecture: from a comprehensive, agent-based system designed to maximize redundancy and robustness, to a more streamlined and efficient model driven by advances in model fluency and OCR reliability.

We began by establishing a baseline architecture that could respond to the multifaceted challenges of digitizing historical documents. This system was conceived not as a linear pipeline but as a constellation of loosely coupled agents, each responsible for a distinct facet of the correction process. The notion of an "agent" is deployed here in its most general and flexible sense—any discrete process, tool, or model that can operate independently yet collaboratively within the broader architecture. This design allowed us to capture the granular complexity of historical variation while providing a modular framework for experimentation.

At its core, the baseline architecture, (See Figure 1), integrated a preprocessing module with multiple OCR systems whose outputs were then synthesized through a large language model acting as a fusion engine. The selection of three OCR as core components of the OCR pipeline reflects a deliberate strategy grounded in complementarity rather than reliance on a single solution. No single OCR system offers consistently optimal performance across the diverse range of document types, languages, and layouts encountered in large-scale digitization projects. Each tool was chosen for its distinct strengths. The OCR systems selected were Surya and OlmOCR together a Gemini 2.0. *OlmOCR*, developed by AllenAI, is a context-aware toolkit for linearising PDF documents into text, integrating visual–textual co-training and layout modelling to tackle noisy scans and typographic variation with high fidelity, CAllenai (2025) . Its adaptive pipeline excels in multilingual archives and early modern prints being particularly effective for complex or historical documents. *Surya* is an open-source OCR framework optimised for 90+ languages, offering robust line-level text detection, layout analysis (tables, images, headers), and reading-order inference—even on degraded or low-resource scans, CibinQuadance (2025). Its lightweight architecture balances accuracy and speed, making it ideal for transnational DH workflows. Finally, *Gemini 2.0* a Google DeepMind's multimodal large language model augments OCR pipelines by jointly processing text and images, enabling advanced post-OCR tasks such as semantic annotation, layout-aware classification, and multilingual reasoning, without replacing dedicated recognition engines, Google (2025). Together, these systems exemplify the convergence of linguistic insight and machine learning in service of scholarly digitisation. Their integration into Digital Humanities pipelines not only enhances transcription accuracy but also supports interpretability and methodological transparency in AI-assisted research. On the other hand, the fusion model used was GPT-4o, and was tasked with reconciling divergent OCR hypotheses and producing a single, coherent transcription. The fusion model operates by employing a form of Chain of Thought prompting, which guides the language model through a structured sequence of reasoning steps. This approach significantly improves its ability to synthesize fragmented or partially erroneous textual inputs, especially when dealing with noisy OCR outputs. In this study, we designed and tested several prompts tailored to this purpose. One representative example is the following:

```
Prompt=
"You are an expert in text correction and OCR error fixing.
Your task is to combine and correct several OCR outputs of the same text.
Here are the texts:

[Insert OCR outputs here]

Instructions:
1. Combine the texts, correcting any OCR errors.
2. Provide only the corrected text, without any additional commentary.
3. Maintain the original structure and formatting.
4. Do not add any new information or explanations.
5. Join any words that have been separated by a hyphen at the end of a line.
If there're blank spaces after the hyphen,
remove them so the two parts of the word get joined correctly.
6. The text is written using archaic Spanish spelling.
7. Maintain all diacritical marks, old-fashioned spellings,
and historical punctuation, such as the use of 'fué' instead of 'fue',
'dió' instead of 'dio', 'ví' instead of 'vi', 'á' instead of 'a' in prepositions.
Do not replace older words or grammatical structures with modern equivalents.
8. Ensure that all words retain their original diacritics,
such as accents (é, á, ó), tildes (ñ), and umlauts (ü), without alteration.
9. Focus on fixing spelling and obvious OCR mistakes.
10. End your response with '===END===' on a new line.

Corrected text:"
```

*All prompts used in this study are available at:
`https://github.com/sros-UNED/pastreader`

This prompt proved particularly effective for aligning semantically equivalent text fragments and filtering out common OCR errors, while simultaneously preserving the historical orthography and structural fidelity of the source material. By embedding explicit constraints and a stepwise reasoning process, the model is not only more accurate in its outputs but also more robust to noise, misrecognition, and orthographic irregularities in historical documents. Finally, the postprocessing stage corrected residual noise while preserving historically plausible forms, informed by typographic norms and diachronic linguistic data. This architecture is further strengthened by a dedicated preprocessing stage aimed at enhancing image quality prior to OCR. Specifically, adaptive thresholding techniques were employed to improve text-background contrast, while Gaussian corrections were applied to mitigate noise and uneven illumination. These image enhancement methods ensure more reliable character recognition, particularly in degraded or visually inconsistent scans, thereby increasing the overall accuracy and stability of the downstream pipeline.

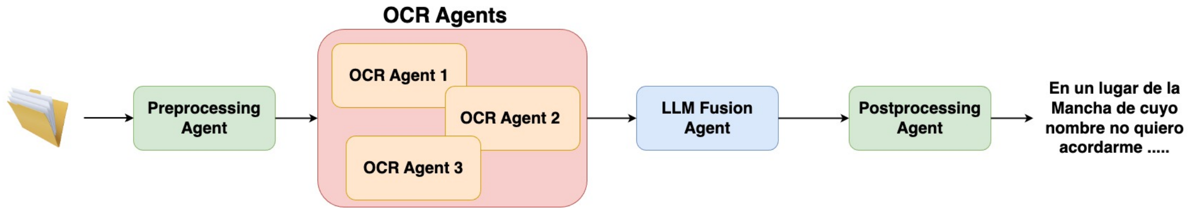


Figure 1: Agentic Architecture

Yet, while this architecture proved effective in principle, its operational overhead prompted us to explore a leaner alternative. Our revised architecture, (See Figure 2), motivated by the empirical results and performance bottlenecks of the baseline, dispenses with multi-engine redundancy and fusion. Instead, it relies on a high-performance OCR system—Google's Gemini 2.5 PRO model, an evolution of the previous uses Gemini 2.0 model—as the sole recognition agent. In this particular implementation, it was observed that even a relatively simple prompt can yield highly effective results when working with advanced multimodal large language models for post-OCR enhancement. The following prompt, for instance, was used with consistent success to extract and normalize historical Spanish text while preserving its linguistic integrity:

```
Prompt=
"Perform OCR (Optical Character Recognition) on this image.
Extract ALL visible text without modernizing or modifying Old Spanish.
Correct spelling and punctuation while preserving the original language and format.
Respond ONLY with the extracted text, without additional comments"
```

Despite its simplicity, this instruction proved sufficient to achieve high-quality transcription results. The model demonstrated strong alignment with the task goals, requiring minimal additional tuning to handle historical orthography and formatting. This highlights the potential of prompt-based architectures to simplify complex processing pipelines through carefully designed language instructions. Also it is not necessary a preprocessing stage, the model makes it straightforward ans its output is passed directly to a postprocessing module, which fulfills the same curatorial role as before, filtering OCR artifacts without imposing anachronistic normalization. This final architecture, while deceptively simple, emerged as remarkably robust, leveraging the maturity of modern OCR engines and the precision of post-hoc linguistic correction to achieve results that rival more complex configurations.

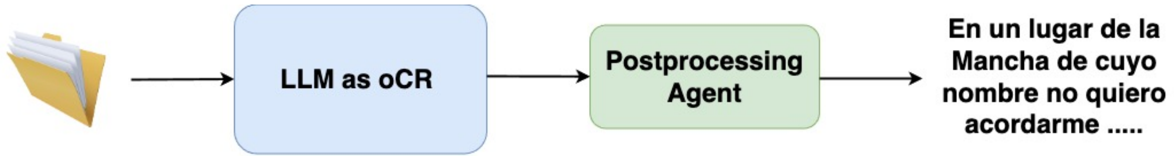

LLM as oCR
Postprocessing Agent
En un lugar de la Mancha de cuyo nombre no quiero acordarme .....

Figure 2: LLM-based Architecture

Finally, both proposed architectures underscore the critical need for a robust postprocessing layer to ensure the usability and fidelity of the OCR-derived textual output. Despite advances in recognition accuracy, raw OCR outputs often include extraneous elements that obscure the primary content and hinder downstream processing. To address this, a comprehensive postprocessing routine was implemented to systematically clean and normalize the extracted text. This includes the removal of institutional attributions (e.g., repeated references to the Biblioteca Nacional de España), typographic artifacts such as copyright symbols (©), pilcrows (¶), or decorative punctuation, and numerical strings often linked to cataloging or pagination. Moreover, the procedure targets frequent OCR errors at the beginning of lines—such as spurious punctuation marks or repeated characters—using regular expressions for precise correction. By trimming superfluous whitespace and blank lines, the postprocessing layer ensures a coherent and structured output, thereby facilitating more reliable linguistic, semantic, or downstream machine learning analyses.

3.1. Dataset Description

The dataset used in this study is derived from the historical press collection digitized by the National Library of Spain (Biblioteca Nacional de España, BNE) and accessible via the *Hemeroteca Digital* platform. As of this writing, the full collection comprises 298 press titles, 88,748 issues, and over 8 million digitized pages. This repository spans publications from the 17th to the 20th centuries and continues to expand through ongoing digitization efforts by the BNE.

For the purpose of development and evaluation, a dedicated **development partition** was curated from this broader corpus. This subset consists of **500 representative pages** selected to ensure a stratified sampling across publication types, time periods, and layout complexities. Each item in the development set includes the following components:

- The original **scanned PDF** page.
- The corresponding **OCR output** (as generated by the legacy OCR system used in Hemeroteca Digital).
- The **manually corrected transcription** serving as ground truth.

These corrected transcriptions were prepared under the BNE's collaborative initiative *ComunidadBNE*, which enables users to contribute to OCR correction on selected publications. Correction projects are chosen based on factors such as historical significance, consultation frequency, and technical feasibility of user-driven annotation.

The OCR quality in the collection varies substantially due to factors like digitization date, scanning technology, page layout complexity, and the physical condition of the originals. The development set, in particular, was assembled to reflect this heterogeneity, thereby ensuring that evaluation metrics are representative of real-world variability in historical press digitization.

The full development dataset is publicly available and can be accessed via the Hemeroteca Digital platform. It forms a critical component of our pipeline design and evaluation, allowing for robust benchmarking of OCR and postprocessing strategies under controlled yet realistic conditions.

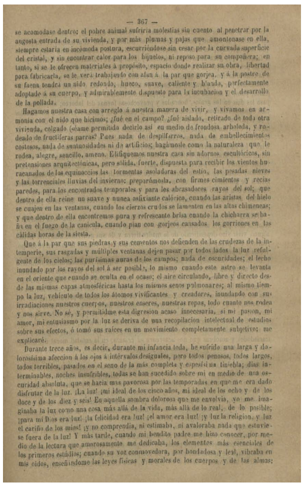
4. Evaluation

To evaluate this system, we adopted a pragmatic sampling strategy grounded in cluster analysis based on k-means algorithm. This process allowed us to simulate a full-corpus evaluation while dramatically reducing computational cost and review time. For this purpose we used the dataset prepared for this task, Montejo-Ráez et al. (2025a). All reference documents used in the clustering process were first inspected to identify the distinct text–image characteristics. With this insight about the structure of the different document in the corpus, we build a feature vector for each document based on de CER and WER metrics compute using three traditional OCR engines. According to the elbow algorithm, the optimal number of cluster was six. Therefore, the corpus, composed of heterogeneous sources spanning three centuries, was algorithmically partitioned into six clusters. For each cluster, the centroid document was selected to identify a single representative document—an exemplar that distilled the dominant characteristics of its group. These six clustering fit with our previous exploratory analysis of the corpus.

Figures 3 and Figure 4 illustrate pages containing only textual lines: the former exhibits the typical yellowing associated with paper aging, while the latter shows minor geometric distortions in the text baselines despite retaining a predominantly white background. Figures 5 and 6 depict pages that combine text with graphical elements and complex structures—titles interleaved with illustrations or multi-column tables—posing challenges for both segmentation and layout analysis. Finally, Figures 7 and 8 present characteristic two-column pages: one cluster features uniform background textures, whereas the other contains varying background tones, necessitating adaptive thresholding and column-detection strategies.

Extending this analysis, each cluster suggests tailored preprocessing and recognition workflows. For the homogeneously aged and white-background text clusters (Figs. 4–8), binarization methods must compensate both for color degradation and slight skew, ensuring accurate line segmentation. The mixed-media clusters (Figs. 5–6) require hybrid layout engines capable of discriminating between textual and graphical regions, often combining connected-component analysis with rule-based heuristics to preserve reading order. Two-column pages with consistent backgrounds (Fig. 7) can leverage fixed grid models to detect column boundaries, whereas those with uneven backgrounds (Fig. 5) benefit from locally adaptive thresholding and morphological filtering to achieve reliable text extraction. Together, these six representative image types form a comprehensive basis for optimizing OCR pipelines and evaluating model robustness across diverse document archetypes.

Each version of the architecture—baseline and final—was applied to these representative texts, enabling us to assess the comparative performance of the systems across a spectrum of historical, typographic and linguistic conditions. Metrics such as Word Error Rate (WER) and Character Error Rate (CER) were calculated to quantify system output.



— 367 —

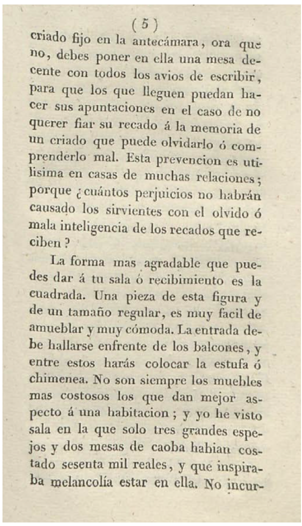
se acomodase dentro; el pobre animal sufriría molestias sin cuento al penetrar por la angosta entrada de su vivienda, y por más plumas y pajas que amontonase en ella, siempre estaría en incómoda postura, escurriéndose sin cesar por la curvada superficie del cristal, y sin encontrar calor para los hijuelos, ni reposo para su compañera; en tanto, si se le ofrecen materiales á propósito, espacio donde realizar su obra, libertad para fabricarla, se le verá trabajando con afan á la par que gorjea, y á la postre de su faena tendrá un nido redondo, hueco, suave, caliente y blando, perfectamente adoptado á su cuerpo, y admirablemente dispuesto para la incubacion y el desarrollo de la pollada.

Hagamos nuestra casa con arreglo á nuestra manera de vivir, y vivamos en armonía con el nido que hicimos; ¿fué en el campo? ¿fué aislado, retirado de toda otra vivienda, colgado (séame permitido decirlo así en medio de frondosa arboleda, y rodeado de fructíferas parras? Pues nada de despilfarros, nada de embellecimientos costosos, nada de suntuosidades ni de artificios; hagámosle como la naturaleza que le rodea, alegre, sencillo, ameno. Edifiquemos nuestra casa sin adornos escultóricos, sin pretensiones arquitectónicas, pero sólida, fuerte, dispuesta para recibir los vientos huracanados de los equinoccios las tormentas asoladoras del estío, las pesadas nieves y las torrenciales lluvias del invierno; preparémosla, con firmes cimientos y recias paredes, para los encontrados temporales y para los abrasadores rayos del sol; que dentro de ella reine un suave y nunca asfixiante calórico, cuando las aristas del hielo se cuajen en las ventanas, cuando los cierzos crudos se lamenten en las altas chimeneas; y que dentro de ella encontremos pura y refrescante brisa cuando la chicharra se baña en el fuego de la canícula, cuando pian con gorjeos cansados los gorriones en las cálidas horas de la siesta.

Que á la par que sus piedras y sus cementos nos defienden de las crudezas de la intemperie, sus rasgadas y múltiples ventanas dejen pasar por todos lados la luz refulgente de los cielos; las purísimas auras de los campos; nada de oscuridades; el lecho inundado por los rayos del sol á ser posible, lo mismo cuando este astro se levanta en el oriente que cuando se oculta en el ocaso; el aire circulando, libre y directo desde las mismas capas atmosféricas hasta los mismos senos pulmonares; al mismo tiempo la luz, vehículo de todos los átomos vivificantes y creadores, inundando con sus irradiaciones nuestros cuerpos, nuestros enseres, nuestras ropas, todo cuanto nos rodea y nos sirve. No sé, y permitidme esta digresion acaso innecesaria, si mi pasion, mi amor, mi entusiasmo por la luz se deriva de una recopilacion intelectual de estudios sobre sus efectos, ó tomó sus raíces en un movimiento completamente subjetivo; me explicaré.

Durante trece años, es decir, durante mi infancia toda, he sufrido una larga y dolorosísima afeccion á los ojos á intérvalos desiguales, pero todos penosos, todos largos, todos terribles, pasados en el seno de la más completa y espesísima tiniebla; dias interminables, noches insufribles, todas se han sucedido sobre mi en medio de una oscuridad absoluta, que se hacia mas pavorosa por las temporadas en que me era dado disfrutar de la luz. ¡La luz! ¡mi ideal de los cinco años, mi ideal de los ocho y de los doce y de los diez y seis! En aquella sombra dolorosa que me envolvia, yo me imaginaba la luz como una cosa más allá de la vida, más allá de lo real, de lo posible; ¡para mí Dios era luz! ¡la felicidad era luz! ¡el amor era luz! ¡y luz la religion, y luz el cariño de los mios! ¡y no comprendia, ni estimaba, ni avaloraba nada que estuviese fuera de la luz! Y más tarde, cuando mi bendito padre me hizo conocer, por medio de la lectura que amorosamente me dedicaba, los elementos más esenciales de los primeros estúdios; cuando su voz conmovedora, por bondadosa y leal, vibraba en mis oidos, enseñándome las leyes físicas y morales de los cuerpos y de las almas;

Figure 3: Text-only page with yellowed background.

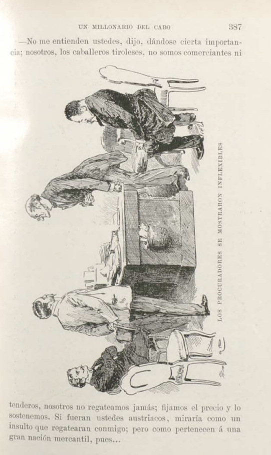


(5)

criado fijo en la antecámara, ora que no, debes poner en ella una mesa decente con todos los avios de escribir, para que los que lleguen puedan hacer sus apuntaciones en el caso de no querer fiar su recado á la memoria de un criado que puede olvidarlo ó comprenderlo mal. Esta prevencion es utilísima en casas de muchas relaciones; porque ¿cuántos perjuicios no habrán causado los sirvientes con el olvido ó mala inteligencia de los recados que reciben ?

La forma mas agradable que puedes dar á tu sala ó recibimiento es la cuadrada. Una pieza de esta figura y de un tamaño regular, es muy fácil de amueblar y muy cómoda. La entrada debe hallarse enfrente de los balcones, y entre estos harás colocar la estufa ó chimenea. No son siempre los muebles mas costosos los que dan mejor aspecto á una habitacion; y yo he visto sala en la que solo tres grandes espejos y dos mesas de caoba habian costado sesenta mil reales, y que inspiraba melancolía estar en ella. No incur-

Figure 4: Text-only page with white background and slight skew.



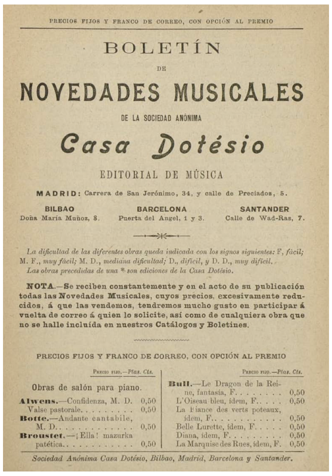
UN MILLONARIO DEL CABO 387

—No me entienden ustedes, dijo, dándose cierta importancia; nosotros, los caballeros tiroleses, no somos comerciantes ni

LOS PROCURADORES SE MOSTRARON INFLEXIBLES

tenderos, nosotros no regateamos jamás; fijamos el precio y lo sostenemos. Si fueran ustedes austriacos, miraría como un insulto que regatearan conmigo; pero como pertenecen á una gran nación mercantil, pues...

Figure 5: Page with text and graphical elements.



PRECIOS FIJOS Y FRANCO DE CORREO, CON OPCIÓN AL PREMIO

BOLETÍN

DE

NOVEDADES MUSICALES

DE LA SOCIEDAD ANÓNIMA

Casa Dotésio

EDITORIAL DE MÚSICA

MADRID: Carrera de San Jerónimo, 34, y calle de Preciados, 5.

BILBAO	**BARCELONA**	**SANTANDER**
Doña María Muñoz, 8.	Puerta del Angel, 1 y 3.	Calle de Wad-Ras, 7.

La dificultad de las diferentes obras queda indicada con los signos siguientes: F, *fácil;* M. F., *muy fácil;* M. D., *mediana dificultad;* D., *difícil, y* D. D., *muy difícil.*

*Las obras precedidas de una * son ediciones de la Casa Dotésio.*

NOTA.—Se reciben constantemente y en el acto de su publicación todas las Novedades Musicales, cuyos precios, excesivamente reducidos, á que las vendemos, tendremos mucho gusto en participar á vuelta de correo á quien lo solicite, así como de cualquiera obra que no se halle incluída en nuestros Catálogos y Boletines.

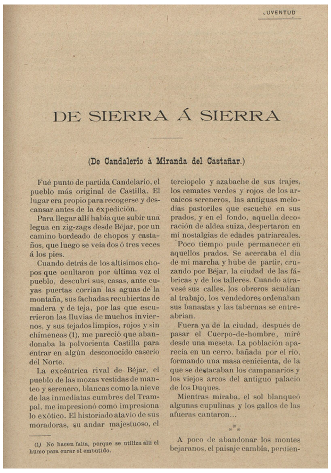
PRECIOS FIJOS Y FRANCO DE CORREO, CON OPCIÓN AL PREMIO

	Precio fijo.—*Ptas. Cts.*		Precio fijo.—*Ptas. Cts.*
Obras de salón para piano.		**Bull.**—Le Dragon de la Reine, fantasía, F.	0,50
Alwens.—Confidenza, M. D.	0,50	L'Oiseau bleu, ídem, F.	0,50
Valse pastorale.	0,50	La Fiance des verts poteaux, ídem, F.	0,50
Botte.—Andante cantabile, M. D.	0,50	Belle Lurette, ídem, F.	0,50
Broustet.—¡Ella! mazurka patética.	0,50	Diana, ídem, F.	0,50
		La Marquise des Rues, ídem, F.	0,50

Sociedad Anónima Casa Dotésio, Bilbao, Madrid, Barcelona y Santander.

Figure 6: Page with multi-column tables and complex layout.

In Table 1, the first three column-pairs—*Gemini 2.0*, *Surya*, and *OlmOCR*—report the *Character Error Rate* (CER) and *Word Error Rate* (WER) immediately after each OCR engine has processed the document images through the dedicated preprocessing pipeline (binarization, deskewing, layout normalization, etc.). These values thus capture the raw recognition performance of each system on cleaned inputs. The fourth column-pair, labeled *GPT-4o*, presents CER and WER measured after a *fusion* step in which OCR outputs are combined and disambiguated by GPT-4o's multimodal reasoning. By leveraging contextual cues across line segments and graphical regions, this fusion reduces insertion, deletion, and substitution errors, yielding consistently lower error rates than any individual OCR source. The *POSTPRO* column-pair shows CER and WER following a post-processing stage.



JUVENTUD

DE SIERRA Á SIERRA

(De Candalerio á Miranda del Castañar.)

Fué punto de partida Candelario, el pueblo más original de Castilla. El lugar era propio para recogerse y descansar antes de la expedición.

Para llegar allí había que subir una legua en zig-zags desde Béjar, por un camino bordeado de chopos y castaños, que luego se veia dos ó tres veces á los pies.

Cuando detrás de los altísimos chopos que ocultaron por última vez el pueblo, descubrí sus, casas, ante cuyas puertas corrían las aguas de la montaña, sus fachadas recubiertas de madera y de teja, por las que escurrieron las lluvias de muchos inviernos, y sus tejados limpios, rojos y sin chimeneas (1), me pareció que abandonaba la polvorienta Castilla para entrar en algún desconocido caserio del Norte.

La excéntrica rival de Béjar, el pueblo de las mozas vestidas de manteo y serenero, blancas como la nieve de las inmediatas cumbres del Trampal, me impresionó como impresiona lo exótico. El historiado atavío de sus moradoras, su andar majestuoso, el terciopelo y azabache de sus trajes, los remates verdes y rojos de los arcaicos sereneros, las antiguas melodías pastoriles que escuché en sus prados, y en el fondo, aquella decoración de aldea suiza, despertaron en mí nostalgias de edades patriarcales.

Poco tiempo pude permanecer en aquellos prados. Se acercaba el día de mi marcha y hube de partir, cruzando por Béjar, la ciudad de las fábricas y de los talleres. Cuándo atravesé sus calles, los obreros acudían al trabajo, los vendedores ordenaban sus banastas y las tabernas se entreabrían.

Fuera ya de la ciudad, después de pasar el Cuerpo-de-hombre, miré desde una meseta. La población aparecía en un cerro, bañada por el río, formando una masa cenicienta, de la que se destacaban los campanarios y los viejos arcos del antiguo palacio de los Duques.

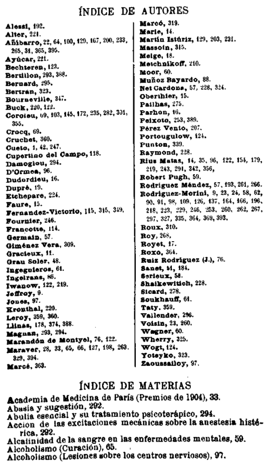
Mientras miraba, el sol blanqueó algunas cupulinas y los gallos de las afueras cantaron...

***

A poco de abandonar los montes bejaranos, el paisaje cambia, perdien-

(1) No hacen falta, porque se utiliza allí el humo para curar el embutido.

Figure 7: Two-column page with yellowed background.



ÍNDICE DE AUTORES

Alessi, 192.
Alter, 221.
Añibarro, 22, 64, 100, 120, 167, 200, 233, 265, 31, 365, 395.
Ayúcar, 221.
Bechteren, 123.
Bertillon, 293, 388.
Bernard, 295.
Bertrán, 323.
Bourneville, 347.
Buck, 220, 122.
Coroleu, 69, 103, 145, 172, 235, 282, 301, 355.
Crocq, 69.
Cruchet, 360.
Cueto, 1, 42, 247.
Cupertino del Campo, 118.
Damoglou, 294.
D'Ormea, 96.
Dudordieu, 16.
Dupré, 19.
Etchepare, 224.
Faure, 15.
Fernandez-Victorio, 115, 315, 349.
Fournier, 246.
Francotte, 114.
Germain, 57.
Giménez Vera, 309.
Gracieux, 11.
Grau Soler, 48.
Ingegnieros, 61.
Ingelrans, 86.
Iwanow, 122, 219.
Joffroy, 9.
Jones, 97.
Kronthal, 220.
Leroy, 359, 360.
Llinas, 178, 374, 388.
Magnan, 293, 294.
Marandón de Montyel, 76, 122.
Maraver, 28, 33, 65, 66, 127, 198, 263. 329, 394.
Marcé, 363.
Marcó, 319.
Marie, 14.
Martín Istúriz, 129, 203, 231.
Massoin, 315.
Meige, 18.
Metchnikoff, 210.
Moor, 60.
Muñoz Bayardo, 88.
Net Cardona, 57, 228, 324.
Oberthier, 15.
Pailhas, 275.
Parhon, 16.
Peixoto, 253, 389.
Pérez Vento, 207.
Portougulow, 124.
Punton, 339.
Raymond, 228.
Rius Matas, 14, 35, 96, 122, 154, 179, 219, 243, 291, 342, 356,
Robert Pugh, 59.
Rodríguez Méndez, 57, 193, 261, 266.
Rodríguez-Morini, 9, 23, 24, 58, 62, 90, 91, 98, 109, 126, 137, 164, 166, 196, 218, 223, 229, 246, 253, 260, 262, 267, 297, 327, 335, 364, 369, 393.
Roux, 310.
Roy, 268.
Royet, 17.
Roxo, 364.
Ruiz Rodriguez (J.), 76.
Sanet, 51, 184.
Serieux, 58.
Shaikewitich, 228.
Sicard, 278.
Soukhauff, 61.
Taty, 359.
Vallender, 296.
Voisin, 23, 260.
Wagner, 60.
Wherry, 325.
Wogt, 124.
Yoteyko, 323.
Zaoussailoy, 97.

ÍNDICE DE MATERIAS

Academia de Medicina de París (Premios de 1904), 33.
Abasia y sugestión, 292.
Abulia esencial y su tratamiento psicoterápico, 294.
Accion de las excitaciones mecánicas sobre la anestesia histérica, 292.
Alcalinidad de la sangre en las enfermedades mentales, 59.
Alcoholismo (Curación), 65.
Alcoholismo (Lesiones sobre los centros nerviosos), 97.

Figure 8: Two-column page with uniform background.

Taken together, these results illustrate how successive stages—preprocessing, multimodal fusion, and linguistic post-processing—each contribute to progressive error reduction demonstrating a clear trajectory of improvement in both character- and word-level accuracy.

$$\underbrace{\text{Preprocessing}}_{\substack{\text{removes}\\ \text{image artifacts}}} \longrightarrow \underbrace{\text{Fusion}}_{\substack{\text{leverages}\\ \text{complementary outputs}}} \longrightarrow \underbrace{\text{Postprocessing}}_{\substack{\text{enforces}\\ \text{grammatical consistency}}} \longrightarrow \begin{cases} \text{CER} < 2\% \\ \text{WER} \approx 4\% - 8\% \end{cases}$$

Table 1
Comparison of Character Error Rate (CER) and Word Error Rate (WER) across OCR systems and preprocessing pipelines for six representative document clusters.

Cluster	**Gemini 2.0**		**Surya**		**OlmOCR**		**GPT-4o**		**POSTPRO**		**Gemini 2.5 Pro Only**	
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
Cluster 1	1,03	3,99	2,74	7,84	2,79	9,02	1,35	5,18	1,25	4,59	1,42	5,17
Cluster 2	3,35	10,73	1,01	4,52	3,04	9,04	1,93	6,21	2,23	7,34	2,03	5,71
Cluster 3	0,28	1,96	12,01	11,76	2,79	1,96	0,28	1,96	0,28	1,96	0,28	1,96
Cluster 4	27,07	29,41	26,1	38,46	22,8	48,42	27,85	30,77	12,63	24,89	4,32	10,64
Cluster 5	2,14	5,51	1,33	4,06	1,48	4,93	1,19	3,19	1,19	3,19	8,32	11,27
Cluster 6	2,46	11,07	2,51	11,07	80,25	82,74	2,15	9,77	2,35	10,42	2,04	9,42
Avg	**4,20**	**7,88**	**5,62**	**10,18**	**12,14**	**18,76**	**3,96**	**7,22**	**2,76**	**7,20**	**2,86**	**6,74**

Finally the fifth column-pair, labeled *Gemini 2.5 Pro Only*, presents CER and WER measures of the second architecture, where we only use a simple prompt and this LLMs for the full process. The results obtained from the second architecture demonstrate performance comparable to the final output of the initial architecture. We would like to highlight that the only case in which the model is outperformed by the first architecture involves a document featuring a title and two-column layout. This discrepancy arises because Gemini extracts the information in a structurally different manner compared to the ground-truth annotations of that document. A simple post-processing step would be sufficient to resolve this issue. Therefore, unlike the latter, the second approach offers significant advantages in terms of cost-efficiency and implementation simplicity. Therefore, models such as Gemini 2.5 Pro rightfully deserve recognition as state-of-the-art performers, combining high accuracy with streamlined deployment. What emerges from this iterative methodological refinement is not merely a more effective pipeline, but a new paradigm for historical OCR correction. The shift from complex fusion to efficient minimalism mirrors a broader transition in the field: one in which redundancy is no longer required to compensate for the limits of AI, but becomes an unnecessary burden in the face of increasingly fluent and historically aware models.

4.1. OCR system selection for evaluation process

In the present study, we adopted a two-tiered approach to OCR-based document analysis, whereby different models were strategically employed across distinct stages of the experimental workflow. Specifically, PaddleOCR, Doctr, and Surya were utilized during the clustering and document selection phase, whereas the evaluation of OCR architectures was conducted using Surya, Gemini 2.0, and OlmOCR. This methodological divergence is grounded in both technical rationale and the differentiated objectives intrinsic to each stage of analysis. During the initial phase—dedicated to document clustering and the identification of key representative texts—we deliberately employed a heterogeneous ensemble of OCR systems. By integrating PaddleOCR, Doctr, and Surya, we sought to harness the complementary biases and strengths of each engine. This ensemble strategy allowed for a richer and more nuanced textual representation of the corpus, thereby enhancing the granularity and fidelity of the clustering process. Such an approach is supported by the principle of complementary redundancy, whereby aggregating outputs from diverse OCR systems mitigates model-specific artifacts and augments the representativeness of selected documents. In this context, OCR outputs were not evaluated per se, but served as intermediary linguistic proxies, enabling the unsupervised grouping of documents along latent textual patterns.

In contrast, the subsequent evaluation phase necessitated a different set of criteria, oriented toward the quantitative and qualitative assessment of OCR performance. Here, the focus shifted to a comparative study of advanced architectures, namely Surya, Gemini 2.0, and OlmOCR. Surya was retained as a baseline to preserve methodological continuity across phases, while Gemini 2.0 and OlmOCR were selected for their capacity to address domain-specific challenges—such as degraded text, historical typography, and script variability. This targeted selection reflects the need for precision and robustness in OCR tasks involving complex or noisy input, particularly within heritage and archival contexts. Moreover, the evaluation was carried out on a curated subset of documents identified during the clustering phase, ensuring that the testbed was both representative and sufficiently challenging.

The use of distinct OCR models across the pipeline also reflects a broader methodological stance: that of functional differentiation. In complex document processing workflows, it is methodologically sound—and increasingly common—to delegate distinct tasks to specialized components. Clustering, by its exploratory and heuristic nature, benefits from model plurality and interpretative breadth; evaluation, by contrast, demands controlled variables and analytic precision. In summary, the bifurcation in model selection reflects a deliberate alignment between methodological purpose and technical affordance. A diverse OCR ensemble facilitated robust document clustering and selection, while the evaluation phase privileged state-of-the-art architectures to benchmark recognition performance under challenging conditions. This stratified approach enabled a more nuanced and effective exploration of OCR capabilities across the document corpus.

5. Task resolution

Following the experiments presented above, and after confirming that the architecture based solely on a large multimodal language model (LLM) achieved the best performance metrics in terms of Character Error Rate (CER) and Word Error Rate (WER), we proceeded to apply this architecture to the text dataset provided by the *PastReaders* task. The officially reported results are as follows, Table 2:

It can be observed that our system outperformed the proposed baseline across all benchmarks. Likewise, it surpassed all participants in the task, demonstrating that multimodal LLMs offer remarkable performance combined with exceptional ease of use. According with the requirements of the task we include the results of CO_2´s emissions of our system, Table 3.

Table 2
Evaluation metrics for the *PastReaders* task dataset

Team	Levenshtein	WER	Norm. Edit	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
OCRTIST	56.3023	0.2344	0.0191	0.8035	0.8849	0.8065	0.8834	0.8843
Baseline	98.4497	0.3725	0.0334	0.6083	0.8244	0.7014	0.8190	0.8237

Table 3
CO_2 emissions' arameters

Parameter	Value
duration	58553.213
emissions	0.159
cpu_energy	0.691
gpu_energy	0.223
ram_energy	0.001
energy_consumed	0.916
cpu_count	32
gpu_count	1
cpu_model	13th Gen Intel(R) Core(TM) i9-13900
gpu_model	NVIDIA RTX 5000 Ada Generation
ram_total_size	62.503

6. Conclusion

This study demonstrates the effectiveness of a tiered, empirically driven approach to OCR pipeline design for historical Spanish documents, leveraging both traditional recognition engines and multimodal large language models. By first establishing a modular, agent-based architecture and subsequently simplifying it through the empirical evaluation of its components, we illustrate a trajectory from redundancy-oriented robustness to streamlined efficiency without sacrificing transcription quality.

Our results confirm that the integration of advanced OCR engines such as Gemini 2.0, alongside fusion via GPT-4o and postprocessing, yields significant improvements in character and word recognition rates across heterogeneous document types. Notably, each stage of the pipeline—preprocessing, fusion, and postprocessing—contributed cumulatively to error reduction, with the final output exhibiting CER below 2% and WER consistently between 4% and 8%.

Through cluster-based sampling and representative document selection, we ensured that system evaluation was both computationally tractable and methodologically representative. This stratified approach uncovered critical layout-dependent challenges, especially in documents with complex structures (e.g., multi-column layouts or mixed graphical-text content), which informed the design of adaptive recognition strategies.

Crucially, our comparative analysis of the second, leaner architecture—based solely on the Gemini 2.5 Pro model—shows that its performance closely matches, and in most cases surpasses, the more complex baseline architecture. The only notable discrepancy, observed in a single two-column layout case, results from a structural misalignment between the model's output and the ground-truth segmentation, a difference that can be resolved with minimal postprocessing. This finding underscores the growing maturity of modern OCR-LM integrations and their capacity to deliver state-of-the-art results with minimal configuration overhead.

Beyond performance, the implications of this work are methodological. The use of distinct OCR systems across pipeline stages—ensemble models for clustering and high-precision engines for evaluation—highlights the value of functional differentiation. By aligning model capabilities with the specific affordances and requirements of each task, we achieved both interpretive depth and evaluative rigor. The results advocate for a modular yet adaptive approach to document processing in digital humanities workflows, one that balances the need for precision with the practicalities of scalability and cost.

Ultimately, our findings position models such as Gemini 2.5 Pro not merely as enhancements to traditional OCR workflows but as viable standalone solutions for historical document transcription—combining accuracy, efficiency, and accessibility. As large language models continue to evolve, their integration into scholarly digitization pipelines promises to redefine both the technical boundaries and methodological assumptions of textual heritage research.

Acknowledgments

This research has been supported by the CLS INFRA project (Grant Agreement ID: 101004984), funded by the European Union. Special thanks to the funding agency for providing the necessary resources to make this work possible.

Declaration on Generative AI

During the preparation of this work, the author(s) used LLM in order to: Develop and evaluate the architectures, grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Online Resources

The sources for this work are available at:

- Pastreaders Code,

The full code and prompts used are included in the repository.

References

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S., and Kaplan, F. (2024). Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study. In Bizzoni, Y., Degaetano-Ortlieb, S., Kazantseva, A., and Szpakowicz, S., editors, *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

CAllenai (2025). OlmOCR. original-date: 2024-09-17T14:53:40Z.

CibinQuadance (2025). surya-OCR. original-date: 2024-02-21T07:43:20Z.

González-Barba, J. Á., Chiruzzo, L., and Jiménez-Zafra, S. M. (2025). Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org.*

Google (2025). Gemini.

Greif, G., Griesshaber, N., and Greif, R. (2025). Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. arXiv:2504.00414 [cs] version: 1.

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H., and Ginis, V. (2025). Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. arXiv:2501.11623 [cs] version: 1.

Manrique-Gómez, L., Montes, T., Rodríguez-Herrera, A., and Manrique, R. (2024). Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 132–139. arXiv:2407.12838 [cs].

Montejo-Ráez, A., Sánchez Nogales, E., Expósito Álvarez, G., Ureña López, A., Martín-Valdivia, M. T., Collado-Montañez, J., Cabrera de Castro, I., Cantero Romero, M. V., García Serrano, A., Ortuño Casanova, R., and Torterolo Orta, Y. A. (2025a). Pastreader 2025. https://doi.org/10.5281/zenodo.15084265. [Data set].

Montejo-Ráez, A., Sánchez-Nogales, E., Expósito-Álvarez, G., Ureña-López, L. A., Martín-Valdivia, M. T., Collado-Montañez, J., Cabrera-de Castro, I., Cantero-Romero, M. V., and Ortuño-Casanova, R. (2025b). Overview of pastreader shared task in iberlef 2025: Transcribing texts from the past. *Procesamiento del Lenguaje Natural*, 75.

Subramani, N., Matton, A., Greaves, M., and Lam, A. (2021). A Survey of Deep Learning Approaches for OCR and Document Understanding. arXiv:2011.13534 [cs].

Yenigün, O. (2025). Fine-Tuning T5 for Grammar Correction: A Step-by-Step Guide.