Fine-Tuning a Compact Multimodal Model on Consumer-Grade Hardware at PastReader 2025

Yanco Amor Torterolo-Orta[1,*,†], Marina Miguez-Lamanuzzi[1,†]

[1]Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Abstract

This paper presents the experiments conducted as part of the IberLEF 2025 shared task *PastReader: Transcribing Texts from the Past*. A compact, open-source multimodal model was fine-tuned on the dataset provided by the organising team. The model was run on consumer-grade hardware, as one of our main goals is to democratise the use of AI while contributing to the preservation and digitisation of historical documents. The results were notable given the hardware limitations. Future work will focus on evaluating other models of similar size and exploring new techniques, with a special emphasis on leveraging the shared task dataset in collaboration with the Biblioteca Nacional de España (BNE).

Keywords

OCR, Historical Documents, Digital Humanities, Multimodal OCR, Fine-tuning, Granite3.2-vision

1. Introduction

Documents from the past have much to say in the present, just as history teaches people so that it does not repeat itself. However, old documents with significant historical and cultural value can be challenging to use computationally, hindering efforts to extract information from them. This is why the shared task *PastReader: Transcribing Texts from the Past* [1] is so relevant. This shared task was organized as part of the 2025 edition of the Iberian Languages Evaluation Forum (IberLEF 2025) [2].

This paper explores our participation in this competition. Our teams is composed of a philologist and a computational linguist. Previous related work includes the organisation of *The Financial Document Causality Detection Shared Task (FinCausal 2025)* [3], and the best-scoring participation at the shared task *FinancES 2023: Financial Targeted Sentiment Analysis in Spanish* [4].

Historically, OCR has always been one of the main topics of research and development given its real-world applications. In fact, it is increasingly drawing attention thanks to recent advancements in multimodal models, both open-source and proprietary. *Llava* [5] is a well-known multimodal model, but some open-source models that provide a multimodal version include Meta's *Llama3.2-vision* [6] or Google's *Gemma3-vision* [7]. On the proprietary side, OpenAI offer several multimodal models like *GPT-4-turbo*, *GPT-o4-mini* or *GPT-o3*[1]. Currently, there is a vast number of research leveraging multimodal models for OCR. One example is [8], with an open-source toolkit available for people to fine-tune their models on English OCR tasks. They provide both the dataset used, *olmOCR-mix-0225*, and their fine-tuned model, *olmOCR-7B-0225-preview*, which is based on *Qwen2-VL-7B-Instruct* [9]. Another example can be found in [10], which presents a multimodal model for generating captions for archaeological material that is often poorly referenced or lacks appropriate textual descriptions. It is worth noting that the emergence of the Digital Humanities and the growing awareness of the importance of preserving and digitising analogue data for future use make this task particularly relevant [11]. Our approach to this shared task consists of fine-tuning Granite, an open-source model developed by IBM [12].

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

ytorterolo@lsi.uned.es (Y. A. Torterolo-Orta); mmiguez93@alumno.uned.es (M. Miguez-Lamanuzzi)

0000-0002-3688-3293 (Y. A. Torterolo-Orta); 0000-0002-1941-8031 (M. Miguez-Lamanuzzi)

[1]https://platform.openai.com/docs/models/compare

The rest of this work is structured as follows: section 2 contextualises the shared task and introduces our approach; while section 3 focuses on an analysis of the dataset, highlighting certain aspects that hinder model performance. Section 4 describes our Granite-based approach in detail, and section 5 provides an in-depth analysis of the results. Some conclusions are drawn at the end in section 6, offering insights into the task and directions for future work.

2. Task description and approach

PastReader 2025 consists of applying OCR on PDF files with limited quality. Two different, yet related, tasks were proposed. On the one hand, OCR outputs from said PDF files are provided. These outputs are in plain text and purposely exhibit a suboptimal quality. Hence, the task 1 is to provide clean versions of these texts. The clean texts are meant to align with the ground truth. On the second hand, the task 2 requires developing an end-to-end OCR system. PDF files are the intended input, ignoring the proposed OCR text outputs from task 1. From the PDF files, new text outputs are expected to be provided directly. The main challenge lies in the dataset quality and the variety of the sources.

As anticipated, our participation consists of fine-tuning *Granite3.2-vision:2b* [13] on the provided dataset. This model was selected over other open-source multimodal models for two main reasons. The first one is that we value the democratisation of LLMs and being able to fine-tune small models on consumer-grade hardware. This is a small model, with only 2 billion parameters, which raised the chances of it fitting in the 16gb of VRAM of the Nvidia RTX 5080 employed for this matter. The second reason is the high performance this model offers despite its size.

According to IBM[2], this model focuses on being especially performant for enterprises. It has been fine-tuned on IBM's DocFM dataset. It is a "large instruction tuning dataset" consisting of high-quality enterprise data. Basically, it prioritises visual document understanding with both image and text, which encompasses document characteristics such as layouts, fonts, charts, etc. Conversely, other models are mainly trained with natural images, allegedly yielding worse results than IBM's model. In fact, based on IBM's claims, this model rivals larger models in benchmarks such as DocVQA and ChartQA. Granite stands out in document understanding and multimodal retrieval-augmented generation (RAG). The remarkable ability of multimodal models to "understand" images and answer questions about said images makes them a perfect choice for OCR. Therefore, it was considered a suitable choice for this shared task.

Given time constraints and the end-to-end nature of Granite's OCR system, we decided to participate in Task 2 only.

3. Datasets analysis

The dataset [14] provided consisted of historical documents sourced from various archives and repositories. As a result, it was highly heterogeneous in nature, including variations in font styles, print quality, background colouration, presence of handwritten marginalia, and embedded visual elements such as stamps or illustrations. This diversity closely reflects the real-world conditions under which digitisation efforts must operate, but it also presents a significant challenge for OCR systems that are typically trained on more standardised inputs.

These inconsistencies in the visual presentation of the documents have a direct impact on OCR performance. Differences in font type and noise introduced by paper degradation or scanning artefacts can significantly increase the character error rate (CER) and word error rate (WER). For example, documents with faded ink or coloured backgrounds may result in missed characters or misclassification, particularly when the contrast between text and background is insufficient.

The generalisation ability of a model trained on one subset of documents may decline when faced with very different styles from another subset. This is especially problematic in historical corpora, where typographic standards and orthographic conventions were far from unified.

[2] https://www.ibm.com/new/announcements/ibm-granite-3-2-open-source-reasoning-and-vision

Granite's broad contextual understanding and vision-language architecture offer promise for interpreting complex layouts and mixed media. However, like any foundation model, its effectiveness is highly dependent on how similar the dataset is to its pre-training data. The wide variation in the dataset can either help by diversifying test conditions or hinder performance if the model has not seen enough similar samples during training. These results underline the importance of dataset consistency or adaptive fine-tuning when deploying OCR systems in heterogeneous historical collections.

3.1. Some interesting aspects of the Dataset

The provided dataset is composed by isolated pages of 8 periodical publications from the 19th and 20th centuries in Spanish language. The topics they cover are very varied, assorted from cultural and general news magazines (*Juventud*, *El Español*), to satirical and humorous magazines (*El Duende Satírico del día*), chronicles of spiritualism (*La luz del porvenir*), serial publications of narrative fiction and poetry (*Revista nueva*, *La patria de Cervantes*) or even newspapers about fashion and feminine customs (*Periódico de las damas*), as well as scientific publications on medical issues (*Revista frenopática española*). As observed, these press pages differ significantly in their goals, themes, and formats.

The following are examples of pages to highlight several complexities that models should be designed to handle. A dedicated model should be trained with these and other common challenges found in historical Spanish texts in mind, as contemporary texts are not a suitable substitute. However, due to time constraints, our team has not yet developed such a model, as creating a dataset with a sufficient number of representative examples remains pending.

Figure 1: File 9062

Figure 2: File 9171

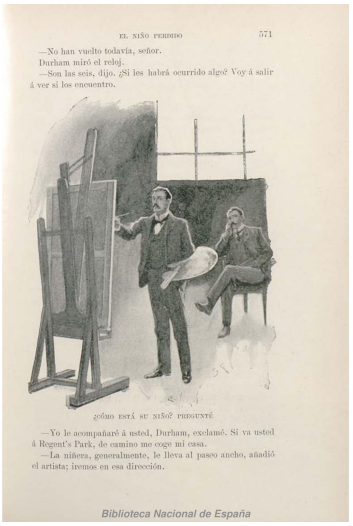

Figure 3: File 8963

First of all we can see that many of the examined documents have in common that they combine images and text on the page, for the most part narrative (novels, short stories), where the image is simply a complement to the narration, to provide it with greater expressive vividness. However, in other cases, there are documents where the image is the main element, and the text is simply a complement to it. For example, this is the case of a page from a fashion magazine (9062, Figure 1), where the text associated with the image is the description of the outfit; in other cases, it is the portrait of a phrenologist doctor (9171, Figure 2), where the associated text is just the doctor's name and their speciality. Thus, it is necessary to take into account that the hierarchy of information is reversed in these two cases, with respect to the informative importance that the images have in relation to the text.

Also, in many of the examined pages, we find that the image, and a brief text associated with it in small capitals, interrupts the transcription of the sentences of the narrative texts (8963, Figure 3). In fact, paratextual elements such as image captions are expected to disrupt the typical linear reading order followed by OCR systems—from top to bottom, left to right. Another aspect that can hinder the OCR is

the presence of typographical ornamentation, like small and decorated capitals. This phenomenon can be observed in the examples above. This can potentially lead the models to misinterpret some letters.

In order to obtain a good training of AI models and reliable results, we consider that it is essential that the dataset is clean and well organised, as well as the application of systematic transcription rules based on coherent choices from a philological and palaeographic point of view. For this reason, we are preparing annotation and transcription guidelines based on the issues detected in the documents of this training dataset for future work.

4. Fine-tuning Granite3.2-vision:2b

The remarkable ability of multimodal models to "understand" images and answer questions about said images makes them a perfect choice for OCR. However, without fine-tuning, i.e., only through prompting, the results are less reliable. These models are prone to analyse pictures or answering relatively complex questions about specific aspects of the image. Therefore, fine-tuning for OCR is highly advisable, mainly to generate the expected transcription exclusively, with no further explanations. Providing a benchmark for *Granite3.2-vision:2b* using only prompting was our intention, but due to time constraints it was eventually impossible. However, despite inconsistencies with format and several issues, initial testing hinted promising results, especially with bigger models like *llama3.2-vision:11b*. Without fine-tuning, *Granite3.2-vision:2b* was more prone to wrongly interpret or misspell Spanish words, probably because its training data is in English language.

We would like to give proper acknowledgments to the author of this ipynb [15], as it provided a preliminary template to work with. In this example, Eli Schwartz fine-tunes *Granite3.1-vision:2b*, the previous version of the model we used, using the Geometric Perception dataset [16] from Hugging Face, hence requiring us to adapt it to the dataset of the shared task, among other changes.

In order to fine-tune the model, the first step was to turn the PDF files into image files, in this case PNG files. In general terms, this kind of models use image files, so any pipeline will generally include a conversion at some point. We performed this conversion in advance for convenience, as it saves some time during processing, although at the cost of an important amount of storage.

After conversion, the dataset consists of the image of the document in PNG format, and the expected output in TXT format. However, this model requires a specific input format, similar to an interaction with a chat model. Figure 4 depicts the structure of the dataset after the input format conversion. As it can bee seen, for each example, it mainly requires a **system prompt**, an **image**, a message from the user, which is defined as **user prompt**, and what we expect from the model: the **ocr text**. The user prompt is: "Please perform OCR on this Spanish document." This is the structure of the dataset, which is passed to the model with the values of the parameters according to each example.

DATASET STRUCTURE

```
chat = [
    "role": "system", "content": ["type": "text", "text": SYSTEM_PROMPT],
    "role": "user", "content": [
        "type": "image", "mage": image,
        "type": "ext", "ext": USER_PROMPT
    ],
    "role": "assistant", "content": ["type": "text", "text": ocr_text],
]
examples.append(chat)
```

Figure 4: Dataset structure required by Granite3.2-vision:2b.

As mentioned, it also requires a system prompt. It is more detailed and contains precise instructions of what is expected from the model. The system prompt used can be seen in Figure 5 below. It instructs

the model not to invent information or modify the text, while trying to obtain the raw text with no further additions.

SYSTEM_PROMPT

```
You are an OCR expert specialised in Spanish documents.
You are analysing an old book scan with potentially low quality.

INSTRUCTIONS:

    Extract ALL text exactly as it appears.

    Do not correct, interpret or modify the text in any way.

    Return ONLY the raw text, without any additional comments or formatting.

    Do not invent content not present in the image.

The output must be EXACTLY the recognised text, without adding anything else.
```

Figure 5: Prompt for OCR with Spanish documents.

It is worth noting that a challenging aspect was the use of RAM memory for loading the dataset image files as pixel values. The employed gaming pc features 32gb of RAM memory. Despite the fact that this is not a low amount of RAM for consumer-grade hardware, it falls short when attempting to process all the images at full resolution from the dataset, running out of memory. Said resolution could vary depending on the file, within the ranges of 1000–1500 pixels by 2000–3000 pixels. Even at a resolution close to 827x1169 pixels, the same error persisted. Considering that *Granite3.2-vision:2b*'s processor scans images by cropping them and analysing areas of 384x384 pixels, we reduced the resolution of the dataset images to a similar size: 414x585 pixels. Using multiples of the crop area dimensions might have been more efficient. It is worth mentioning that, since the images in the dataset have slightly different aspect ratios, some additional measures had to be taken to ensure all images had the same dimensions for fine-tuning. This posed the challenge of potential information loss when resizing them to a fixed resolution given that this process usually crops the images. To address this, each image was resized while preserving its original aspect ratio, ensuring it fitted within the 414×585 pixel target. The remaining space was then filled with padding to reach the exact required dimensions. There are less hardware-demanding alternatives to loading images as pixel values, such as including image paths in the dataset instead of the actual image data and loading them during fine-tuning, or converting the pixel values to tensors ahead of time for later use. However, these options were not explored in this paper.

With respect to quantisation to reduce memory consumption, it is a cornerstone for fine-tuning on consumer-grade hardware. The model was fine-tuned using QLoRA (Quantized Low-Rank Adapter) [17], an approach that combines 4-bit quantisation with parameter-efficient fine-tuning based on LoRA (Low-Rank Adaptation of Large Language Models) [18] adapters. The base model was loaded in 4-bit precision using the NF4 (Normalised Float 4) quantisation scheme, along with double quantisation and bfloat16 computation to optimise numerical stability and performance. Specific modules, such as the vision tower and output head, were excluded from quantisation to preserve their full precision. Lightweight LoRA adapters were injected into the projection layers of the language model, and only these adapters were updated during training. This configuration was fundamental for fitting the model within the RTX 5080's 16gb of VRAM, although it trades off some performance.

Regarding hyperparameters, this aspect remained largely unexplored, as time limitations prevented any meaningful experimentation. The configuration was mostly based on default or initial guesses, with minimal testing. Despite the use of quantisation strategies, special care was still required. Since

the task involves images, VRAM usage was inherently higher. The hyperparameters used are shown in Table 1. As observed, the per-device batch size is set to a minimal value of 1, which is partially mitigated by using 8 gradient accumulation steps. This effectively simulates a larger batch size without exceeding memory limits. In addition, `bfloat16` precision is used alongside gradient checkpointing to further reduce memory usage during training. The fused `adamw_torch_fused` optimiser was selected to maximise computational efficiency. A conservative learning rate of `1e-4` and a weight decay of 0.01 were also applied to ensure stable fine-tuning. Finally, checkpoint saving was limited to the most recent state to save disk space and simplify checkpoint management.

Hyperparameter	Value
Number of training epochs	1
Per-device batch size	1
Gradient accumulation steps	8
Warmup steps	10
Learning rate	`1e-4`
Weight decay	0.01
Logging steps	10
Save strategy	`steps`
Save steps	20
Save total limit	1
Optimizer	adamw_torch_fused
bfloat16 precision	True
Remove unused dataset columns	False
Gradient checkpointing	True
Dataset text field	""
Skip dataset preparation	True

Table 1
Hyperparameters used during fine-tuning.

Even with all the measures and strategies taken to fine-tune *Granite3.2-vision:2b*, it narrowly fitted in the VRAM. The training dataset, comprising both the development and training sets, was randomly split into two parts: 90% for training and 10% for testing. After successfully fine-tuning on the training dataset, the model was used for inference on the final test dataset, made available later by the organisation. Apart from requiring the same dataset format adaptation into a chat format, there was no other significant step worth mentioning. This Granite approach resulted in **GRESEL2_run1**.

5. Analysis of results

5.1. Quantitative analysis

Table 2 offers a chart with the results of the baseline, the winning run and our run. Originally, six different metrics were planned to be implemented in the shared task: Word Error Rate (WER), Sentence Error Rate (SER), Levenshtein Distance, Normalised Edit Distance (NED), BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). However, SER was apparently disregarded, whereas ROUGE was extended to include four variants—ROUGE1, ROUGE2, ROUGEL and ROUGELSUM—, and both WER and Levenshtein Distance were preserved, totalling eight metrics. Consequently, the OCR evaluation relies on a diverse set of complementary metrics that capture different dimensions of performance.

Moreover, literal accuracy is assessed through character-level metrics such as Levenshtein Distance and NED, which count the number of required edit operations; and WER, which quantifies lexical mismatches. Therefore, these metrics favour exact textual reproduction, where lower values indicate better performance. In contrast, semantic quality is measured using BLEU, which evaluates n-gram overlap; and the ROUGE family of metrics: ROUGE-1 focuses on term-level recall, ROUGE-2 captures

TEAM	LEVENSHTEIN	WER	NED	BLEU	ROUGE1	ROUGE2	ROUGEL	ROUGELSUM
OCRTITS_run1	56.3023	0.2344	0.0191	0.8035	0.8849	0.8065	0.8834	0.8843
GRESEL2_run1	97.2399	0.2643	0.0330	0.6890	0.8841	0.8049	0.8806	0.8837
BASELINE	98.4497	0.3725	0.0334	0.6083	0.8244	0.7014	0.8190	0.8237

Table 2
Official leaderboard.

short-range contextual relationships, and ROUGE-L/ROUGE-LSum assess overall discourse coherence. In these metrics, higher values indicate better preservation of linguistic meaning, even when character-level differences are present. Together, these metrics provide a holistic view of system performance, from raw text fidelity to higher-level comprehension. Figure 6 provides a more visual insight into the results.

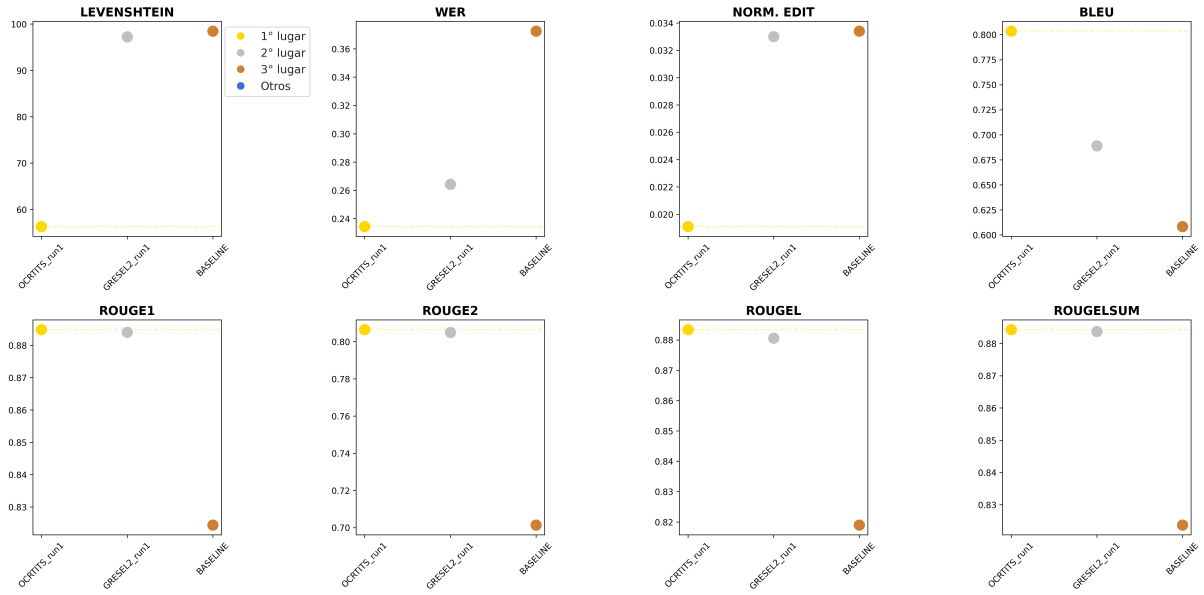

Figure 6: Scattered plots for each metric sorted by best scores.

Table 2 reveals a decent performance in our run. Our fine-tuned *Granite3.2-vision:2b* (**GRESEL2_run1**) achieved a solid second place in five of the eight evaluation metrics, including WER and all ROUGE variants. In fact, the differences between Granite and the top-performing system in these metrics are marginal: ROUGE1 (0.8841 vs. 0.8849), ROUGE2 (0.8049 vs. 0.8065), ROUGE-L (0.8806 vs. 0.8834), ROUGE-LSUM (0.8837 vs. 0.8843), and WER (0.2643 vs. 0.2344). It also ranked third in BLEU. However, Granite underperformed in character-level precision, ranking fourth in both Levenshtein Distance and NED. These results suggest that the model demonstrates excellent semantic understanding and word-level alignment with the ground truth, but has room for improvement in literal character accuracy. Given the compact size of the model and the consumer-grade hardware constraints noted earlier, the performance achieved by *Granite3.2-vision:2b* is highly notable.

Regarding emissions, Figure 7 presents a comparative analysis of computational costs. The left chart uses a logarithmic scale to visualize Granite's fine-tuning and inference metrics during a simulated 10-hour execution window, enabling direct comparison of energy consumption (kWh) and CO_2 emissions (kg) despite their orders-of-magnitude differences. The right chart similarly uses logarithmic scaling to display per-example efficiency metrics across the 3,000-sample dataset. These charts highlight a dramatic difference between the fine-tuning process and the inference process in Granite's run.

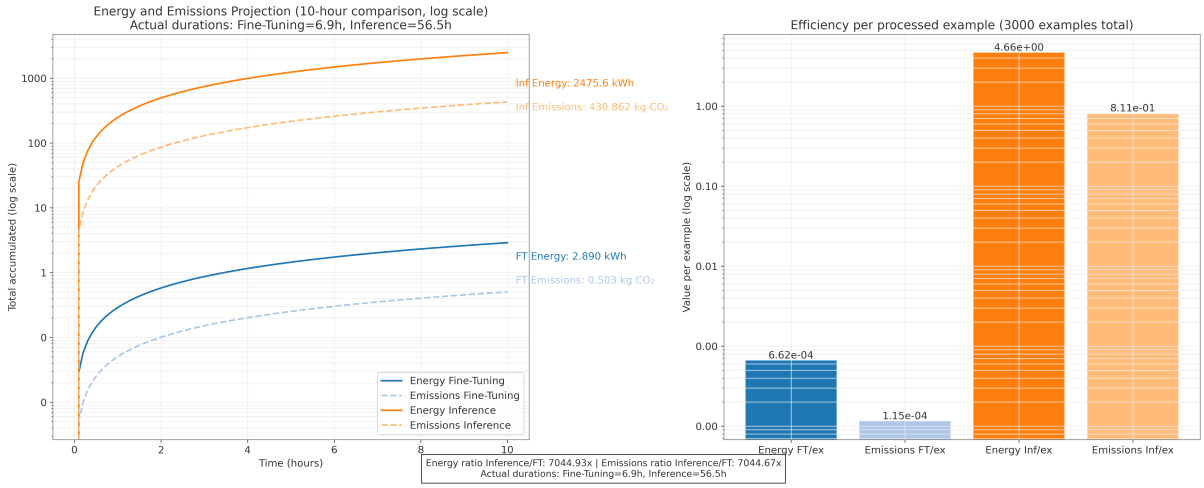

Figure 7: Energy and emission estimations: Granite3.2-vision:2b.

5.2. Qualitative insights

Upon inspecting the output generated by our fine-tuned model, Granite, it was quite satisfactory overall, but some patterns were identified. For instance, when a word is split across lines by a hyphen at the end of a line, the model joins the parts, removes the hyphen, and places the line break after the end of the reconstructed word. The model also misinterprets certain letters and words, particularly those with diacritical marks. This might be due to the fact that *Granite3.2-vision:2b* was not specifically pre-trained on Spanish texts. In general, when the model is unable to recognise a difficult area, it tends to hallucinate content. It also tends to "correct" misspelled words found in the original, even when those words reflect historical variants of Spanish used at the time. These tendencies contribute to lower character-level accuracy, which aligns with the poor results observed in the Levenshtein Distance and NED metrics. On a positive note, the model consistently refrains from generating output when the original file is blank.

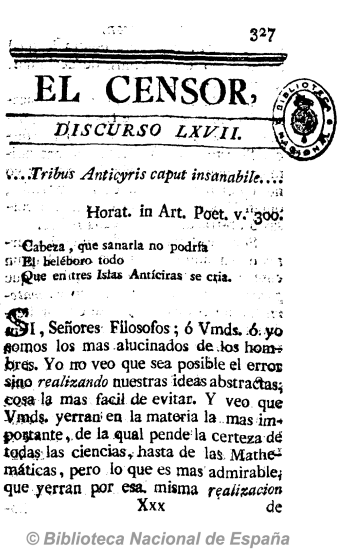

Figure 8: File 2466

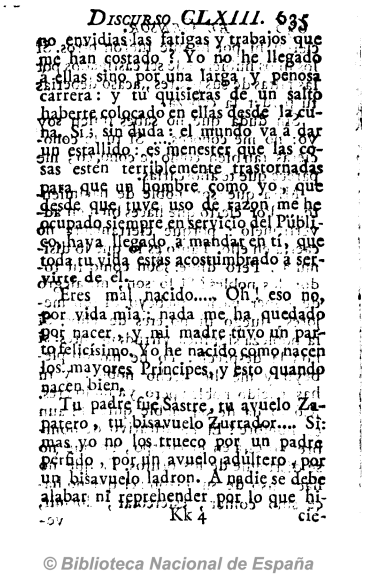

Figure 9: File 410

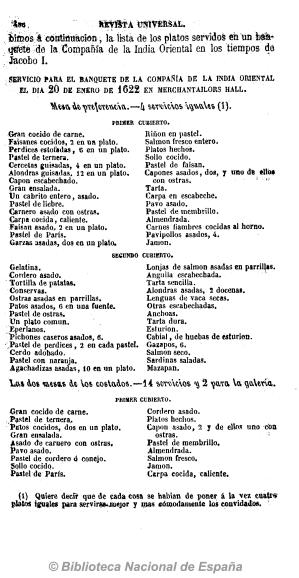

Figure 10: File 1506

It is worth showcasing some challenging examples from the test set. The first example, shown in Figure 8 (file 2466), presents two particularly interesting elements: the stamp featuring the words

BIBLIOTECA NACIONAL (Spanish National Library), and the drop cap in the letter "S." Regarding the words in the stamp, they were ignored by the model, whereas the drop cap was correctly recognised. A word misspelled in historical Spanish, "Mathe-máticas" (Maths), was merged and corrected to *Matemáticas*. In general, the model is more prone to confuse letters or words when the area is blurry or stained.

Regarding the second example, shown in Figure 9, the image might initially appear extremely challenging, as the original physical page seems to have been stained by ink from the facing page on its right. As a result, it contains horizontally mirrored lines superimposed between the actual lines of text. Surprisingly, despite this potential challenge, the model did not struggle with these artefacts. What hindered the recognition of certain words was the presence of blurred letters or smudged areas, yet the overall transcription is quite acceptable given the circumstances.

The last example, Figure 10 displays a two-column layout. This should not pose a major challenge, as OCR models are increasingly better at handling multi-column formats. However, the transcription of this page was particularly inaccurate. The second (right-hand) column was apparently ignored, while the first (left-hand) column was heavily hallucinated. The page contains a menu, and the model invented several dishes for no apparent reason, even adding fabricated information about portion sizes in some cases. This failure to correctly render the column structure might be due to a lack of sufficient multi-column examples in the training dataset—although this remains a hypothesis.

Finally, a last observation about the dataset. While a more thorough analysis would be required to confirm this, the dev/train sets used for fine-tuning appear to be more diverse in terms of sources than the test set. The latter lacks documents containing images, while the former includes materials with different colour palettes in the paper background. Figures 1, 2, and 3, from the dev/train sets, clearly illustrate this. The test set displays less variety in page types and tends to feature whiter backgrounds, in contrast to the yellowish tones found in the dev/train documents. For example, *El Censor* is one of the predominant sources in the test set (Figure 8). This mismatch between the dev/train and test sets might have negatively affected the model's learning and generalisation, although this remains a hypothesis. Ensuring both representativeness and diversity across dataset splits is crucial for effective fine-tuning.

6. Conclusions

This paper has explored our participation in this challenging OCR shared task. Despite the difficulties imposed by the nature and diversity of the dataset, our approach has achieved satisfactory results compared to the baseline. Naturally, some errors commonly associated with AI models were observed, such as the tendency to hallucinate words when the original text is illegible, or to "correct" misspelled words found in the source. We have provided a thorough analysis of the dataset from the competition and offered some ideas about how to fine-tune small multimodal models for OCR tasks and what the expected results are. It is worth highlighting the notable performance of *Granite3.2-vision:2b*, mainly in semantic and word-selection metrics. Although consumer-grade hardware constraints forced the use of a smaller variant and several measures that downgraded fine-tuning quality, it excelled in performance. Overall, this experience has enabled us to learn from our experiments and share our insights into this OCR task.

In future work, we will explore the implementation of a dataset variant from this shared task, but with transcription guidelines focusing on philological and palaeographic criteria, as anticipated. Our hypothesis is that this dataset variant will improve the model's learning. Furthermore, other similarly sized multimodal models will be employed to compare performance between models. Additionally, larger models will be explored, and full-quality fine-tuning will be applied without downgrading the quality of the data through the use of more capable hardware and cloud computing at our disposal. We expect to further enhance our results in OCR tasks by exploring new approaches and ideas.

Acknowledgments

This work is framed under the Spanish National Project GRESEL: UAM (PID2023-151280OB-C21). It has also been partially funded by the PTA2023-023812-I grant, awarded to Yanco Amor Torterolo Orta by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o and Deepseek-V3 in order to: Grammar and spelling check, paraphrase and reword. Besides, Microsoft's Copilot was also used in order to: Formatting assistance (latex commands, image labelling and table creation). Further, the authors used the first two models for figures 6 and 7 in order to: Generate charts based on data. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

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

[2] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, 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, 2025.

[3] A. Moreno-Sandoval, J. Porta, B. Carbajo-Coronado, Y. Torterolo, D. Samy, The financial document causality detection shared task (FinCausal 2025), in: C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, H.-H. Chen (Eds.), Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 214–221. URL: https://aclanthology.org/2025.finnlp-1.21/.

[4] J. Porta-Zamorano, Y. Torterolo-Orta, A. Moreno-Sandoval, LLI-UAM Team at FinancES 2023: Noise, Data Augmentation and Hallucinations, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS, Jaén, Spain, 2023, pp. 1–11. URL: https://ceur-ws.org/Vol-3496/finances-paper1.pdf.

[5] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485. arXiv:2304.08485.

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, (...), Z. Ma, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[7] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, (...), L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[8] J. Poznanski, J. Borchardt, J. Dunkelberger, R. Huff, D. Lin, A. Rangapur, C. Wilhelm, K. Lo, L. Soldaini, olmocr: Unlocking trillions of tokens in pdfs with vision language models, 2025. URL: https://arxiv.org/abs/2502.18443. arXiv:2502.18443.

[9] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang,

M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[10] E. Garcia-Arias, A. Garcia-Serrano, Creación de un modelo de descripciones de imágenes especializado en arqueología griega, Procesamiento del Lenguaje Natural 75 (2025). In press.

[11] A. Garcia Serrano, A. Menta Garuz, La inteligencia artificial en las humanidades digitales: dos experiencias con corpus digitales, Revista de Humanidades Digitales 7 (2022) 19–39. URL: https://revistas.uned.es/index.php/RHD/article/view/30928. doi:10.5944/rhd.vol.7.2022.30928.

[12] M. Mishra, M. Stallone, G. Zhang, Y. Shen, A. Prasad, A. M. Soria, M. Merler, P. Selvam, S. Surendran, S. Singh, M. Sethi, X.-H. Dang, P. Li, K.-L. Wu, S. Zawad, A. Coleman, M. White, M. Lewis, R. Pavuluri, (...), R. Panda, Granite code models: A family of open foundation models for code intelligence, 2024. URL: https://arxiv.org/abs/2405.04324. arXiv:2405.04324.

[13] G. V. Team, L. Karlinsky, A. Arbelle, A. Daniels, A. Nassar, A. Alfassi, B. Wu, E. Schwartz, D. Joshi, J. Kondic, N. Shabtay, P. Li, R. Herzig, S. Abedin, S. Perek, S. Harary, U. Barzelay, A. R. Goldfarb, A. Oliva, B. Wieles, (...), R. Feris, Granite vision: a lightweight, open-source multimodal model for enterprise intelligence, 2025. URL: https://arxiv.org/abs/2502.09927. arXiv:2502.09927.

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

[15] E. Schwartz, Fine-tuning granite vision with trl and peft (lora), https://colab.research.google.com/github/huggingface/cookbook/blob/main/notebooks/en/fine_tuning_granite_vision_sft_trl.ipynb, 2024. URL: https://colab.research.google.com/github/huggingface/cookbook/blob/main/notebooks/en/fine_tuning_granite_vision_sft_trl.ipynb, accessed: 2025-05-08.

[16] J. Zhang, O. Liu, T. Yu, J. Hu, W. Neiswanger, Euclid: Supercharging multimodal llms with synthetic high-fidelity visual descriptions, arXiv preprint arXiv:2412.08737 (2024).

[17] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.