

Handwriting Recognition in Historical Documents with Multimodal LLM*

Abstract

There is an immense quantity of historical and cultural documentation that exists only as handwritten manuscripts. At the same time, performing OCR across scripts and different handwriting styles has proven to be an enormously difficult problem relative to the process of digitizing print. While recent Transformer based models have achieved relatively strong performance, they rely heavily on manually transcribed training data and have difficulty generalizing across writers. Multimodal LLM, such as GPT-4v and Gemini, have demonstrated effectiveness in performing OCR and computer vision tasks with few shot prompting. In this paper, I evaluate the accuracy of handwritten document transcriptions generated by Gemini against the current state of the art Transformer based methods.

Keywords

Optical Character Recognition, Multimodal Language Models, Cultural Preservation, Mass digitization, Handwriting Recognition

1. Introduction

The vast majority of historical documents exist only in manuscript form. Correspondence, personal notes, ledgers, and other unpublished documents provide an enormous source of data that is completely inaccessible to computational text analysis methods. Even for traditional humanists, the lack of keyword search and indexing makes the research process significantly less efficient. Research in these areas requires painstaking and difficult close reading, or the mobilization of huge numbers of volunteer transcribers as in [1]. These barriers prove prohibitive in most cases and limit the scope and scale of potential questions. Unlike the easy and open access to digitized print documents provided by HathiTrust and Internet Archive, databases of handwritten historical documents are nearly nonexistent. These absences in digital availability of sources warp the types of questions scholars are able to pursue, even in traditional humanistic pursuits where travel to physical sources could prove prohibitively expensive. Moreover, recognition and understanding of historical handwriting is often difficult even for subject experts.

While OCR on print has achieved extremely high accuracy (outside of non-Latin types, low quality scanning, and unconventional page formatting) since the early 2000s, high accuracy handwriting OCR was essentially impossible until the adoption of image based convolutional neural networks in studies such as ([2]). These methods are not only computationally intensive, but generally require pretraining or fine tuning on annotated subsets of the specific corpus, creating nearly insurmountable programming barriers for non technically proficient researchers or those without annotated data.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Thus, the development of automated digitization of handwritten documents is an important tool for archives and cultural institutions to open the way for a wide range of new historical research projects, using both conventional close reading and computational approaches.

2. Literature Review

Previous attempts to perform the handwriting task, detailed in surveys like [3] and [4], have generally focused on machine vision models, specifically CNN. However, without a underlying language model, the task was difficult, as many letterforms are inconsistent and indistinguishable, especially in cursive script. An underlying language model was necessary to provide support and predictions in cases of visual ambiguity. Initial studies, like [2], utilized a combined CNN - BiLSTM architecture. The current state of the art ([5], [6],) use a combined vision transformer and text transformer model based on the TrOCR model [7]. This combines a vision transformer encoder with a Roberta based text transformer decoder. TrOCR is trained on synthetic handwriting data and then finetuned for specific domain applications. TrOCR represents the state of the art for general handwriting recognition outside of the historical or multilingual context. Both of these methods rely on extensive preprocessing to align text, detect line breaks, and addressing marginalia, again posing significant barriers to users without technical expertise, along with the aforementioned issues with fine tuning on labelled data.

Multimodal LLMs, such as GPT-4v [8], have displayed promising early ability to recognize text and tabular structure in images. Although architecture and training details are unclear, these models seem to directly take image data along with text prompts as input into a text centric Transformer model. Papers such as [9] and [10] have shown, though limited exploratory examples, that GPT-4 is able to effectively transcribe English handwriting with high accuracy as well as adapt to tabular structure and non aligned text, two problems which previous studies required extensive preprocessing to address. However, these papers have not conducted a comprehensive study on the accuracy of transcription on a robust evaluation dataset. Evaluation of handwriting recognition quality in LLM also focuses on contemporary examples of handwriting, making it difficult to draw conclusions about performance in the historical context.

3. Data

A wide range of multilingual corpora have been published as an evaluation set for this problem. [11] describes the image and gold standard transcription pairs presented at the ICDAR conference between 2014 and 2017, including a 1,200 page excerpt from Jeremy Bentham's papers in Early Modern English, a 450 page set of Early Modern German, and a 150 page dataset of cursive modern Latin. As supplemental sets, [12] provides 200 pages of 17th Century Dutch documents. An additional Spanish language dataset, RODRIGO [13] provided line images for training the SoTA models. These documents were chosen because they have already been used to train and evaluate state of the art models and they represent well structured gold standard labelled data.

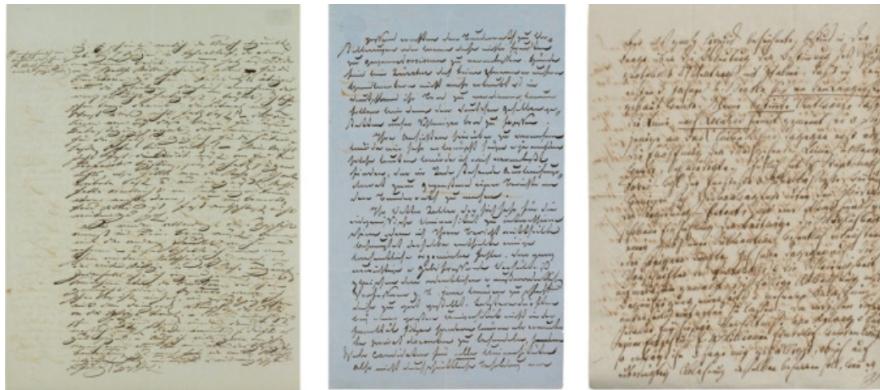


Figure 1: Example data from [11]

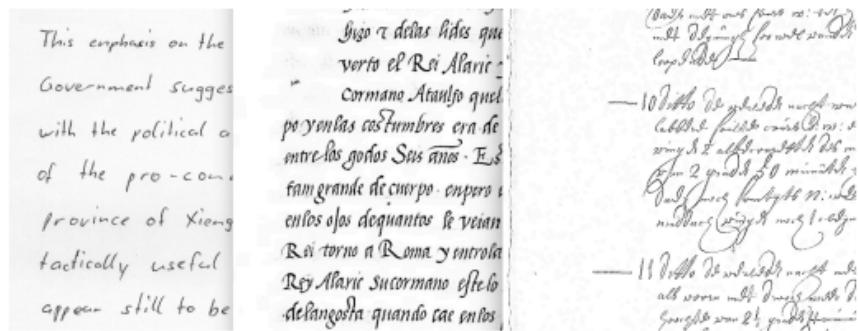


Fig. 1. Example pages IAM, RODRIGO, ScribbleLens

Figure 2: Example data from [12]

4. Methods

4.1. Reproduction

I reimplement a CNN-BiLSTM and a fine-tuned TrOCR architecture based on the SOTA papers. This will be applied to the different letters to evaluate word and character level accuracy to reproduce the results.

For the CNN, I implemented the code at [14] with default hyperparameter settings and roughly 500 epochs of training.

For TrOCR, I fine tuned using the HuggingFace pipeline and TrOCR-Base-Stage-1 with default settings.

Both pre-existing methods were trained or fine tuned using GPU compute on Google Colab. The data was split by language for training and evaluation. The ICDAR 2014 and 2017

Dataset	Line <u>segmented</u>	Language	Era	Page count
ICDAR 2014 and ICDAR 2017 (Bentham papers)	Yes	English	19th Century	1,200
ICDAR 2015 (Ratheskolle)	No	German	17th Century	450
ICDAR 2016	No	German, French, Italian	19th Century	10,000
Rodrigo	Yes	Spanish	16th Century	16,000 lines
ScribbleLens	Yes	Dutch	17th Century	25,000 lines

Figure 3: Metadata Characteristics

Bentham dataset provided 1,200 English samples, ICDAR 2015 provided 450 German documents, and ICDAR 2016 provided 6,000 German and 4,000 French documents. These three languages were selected because they represented the largest collections of languages from specific time periods and corpora available across pre-published data. For each language, I randomly sampled 30 and 500 annotated pages for the training/fine tuning.

4.2. Gemini

The closed source nature of Gemini and other corporate LLMs makes it impossible to fine tune or train with additional data, so the model is applied directly without any modification to fit the specific problems. I selected Gemini, specifically the gemini-pro-vision model, primarily for cost reasons, as the API access is currently free. I evaluate three different prompting strategies. One with just a request to transcribe with no revisions (simple), one providing contextual info on the language and time period of the input image (background), and a final approach where I provide contextual info and ask the model to correct spelling and grammar errors. The simple prompt: "Transcribe the following document exactly with no modifications:". Background prompt: "Transcribe the following document exactly. It is from the X century and in Y language:" Correction Prompt: "Transcribe the following document. It is from the X century and in Y language. Correct any spelling or grammar errors:"

On the initial evaluations, we observe no significant difference between the three prompting strategies. All accuracies were within 1-2 percentage points for each Gemini prompting strategy. This suggests that there is limited impact of providing additional contextual information through prompting beyond the image alone.

It is currently impossible to provide few shot image examples to Gemini because prompts can only contain one image, but as the feature is added I will evaluate further.

5. Results

5.1. Model Performance

Table 1

Character error rate comparison across approaches

Language	CNN-BiLSTM 30 samples	CNN-BiLSTM 500 samples	TrOCR base	TrOCR 30 samples finetune	TrOCR 30 sample fine-tune	TrOCR 500 sample fine-tune	Gemini
English	40.1%	6.8%	62%	31%	7.5%	34%	
French	43.1%	9.8%	81%	43%	10.7%	56%	
German	42.1%	10.1%	92%	53%	9.2%	74%	

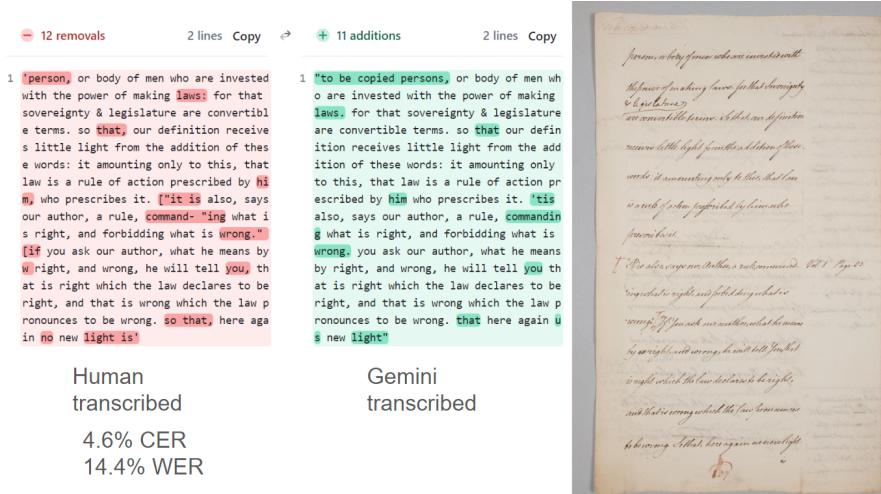


Figure 4: Gemini performance on Bentham document

Figure 4 shows the Gemini output with the prompt: "transcribe the following document. It is in 17th century English by Jeremy Bentham. Correct any spelling errors." and a non pre-processed image provided. We can see in figure 4 that the English language performance of zero shot Gemini transcription is comparable to the SOTA models finetuned on the Bentham corpus. The majority of errors are omissions or incorrect interpretation of punctuation marks (e.g. interpreting a colon as a period or missing brackets added in a different ink color). Figure 5 shows the Gemini output with the prompt: "transcribe the following document. It is in 17th century German." The performance is dramatically worse, and also worse than the SOTA models performance finetuned on 17th century German. This suggests that the English dominated training corpus of Gemini, especially images of handwriting from which it derives emergent handwriting recognition abilities may be impacting performance on certain tasks.

While the overall performance is much worse than fine tuned TrOCR or the CNN, especially for non-English languages, there are some promising signs. First, for English and to a lesser



Figure 5: Gemini performance on German document

extent French and German, the model displays a broad range of performance (Figure 6), with a large number of perfectly transcribed samples, and some high error samples in German resulting from difficult to read text suggesting that the emergent capabilities are present, but more efforts to prevent hallucinations such as lowering temperature or prompting techniques are necessary. When the model fails, it is generally not because of incorrect transcriptions; model errors generally resulted from text generation, rather than letterform recognition. In cases with extremely high inaccuracy, the model generally produced text completely unrelated to the underlying image as a result of hallucination or other model errors (Figure 7). Such hallucinations result from the model’s text generation process, and account for the majority of high error documents, which skews the performance of Gemini. Applying models trained to detect common hallucinations may also reduce the error rate from these occurrences.

6. Discussion

6.1. Recommendations

There are no clear advantages to the state of the art trained models unless a sufficiently large in-domain annotated dataset is provided from the same language and time period. With 30 training samples, the English language performance of both published models is not noticeably superior to zero shot base Gemini. We observe far superior performance with a large number of annotated fine tuning samples, but it is important to remember that this high accuracy is only for a specific corpus, and cannot be easily transferred across languages, time periods, or even specific authors.

Biases in representation of non English languages in the Gemini training dataset makes it much weaker for non-English languages. For these applications, it is still necessary to use trained neural models.

Finetuning with TrOCR, especially with the unique affordances of HuggingFace for simplified

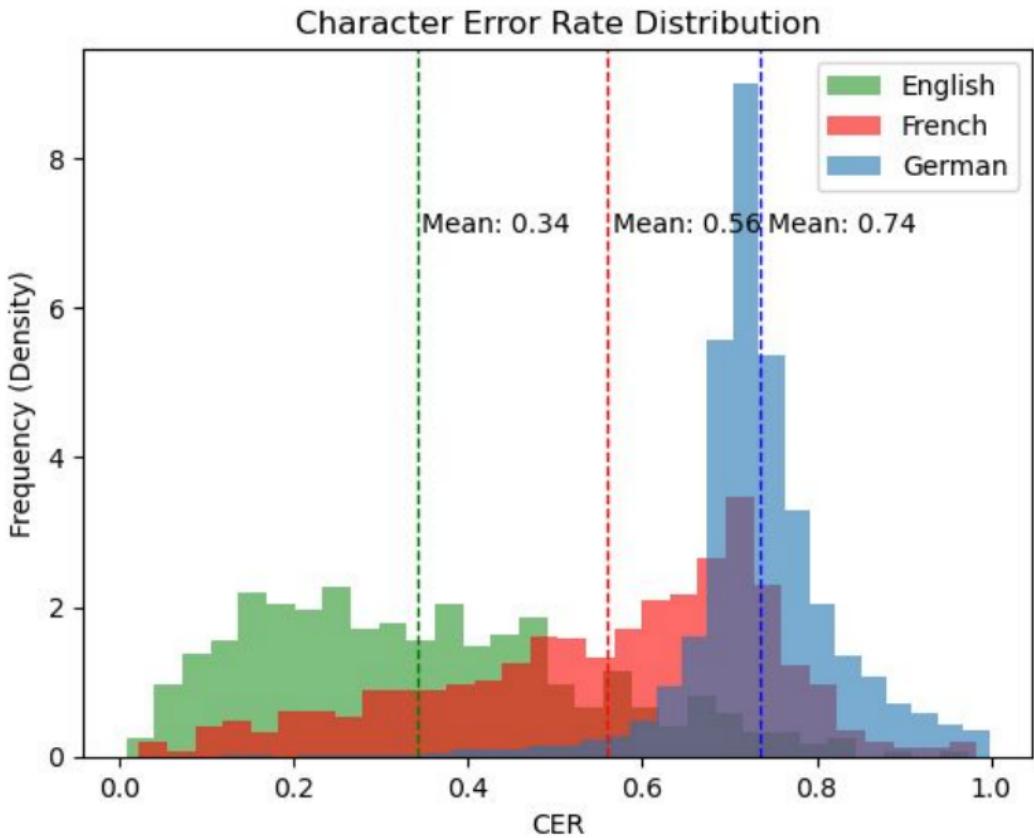


Figure 6: Error Rate Distribution of Gemini

training and model dissemination, appears to offer more cost effective and efficient performance relative to RCNNs. The fine tuning process is both simpler to implement with fewer potential pitfalls versus the raw python code of the RCNN model. At both small and large fine tuning thresholds, TrOCR achieves comparable or superior performance to the RCNN model with a much more streamlined training process.

However, for projects without access to extensive computational resources and/or access to labor for manual training set annotation, Gemini and other multimodal LLMs offers a potentially simpler alternative. Gemini is more accurate at transcribing English language sources than state of the art models with no training or small training sets. On non-English languages, trained models are still necessary.

It is important to also consider the infrastructure necessary to support digitization approaches, and not just the raw accuracy rate. Gemini also allows for programmatically simple and free access (Figure 8), in contrast to previous methods, which rely on access to GPU resources, previous experience with fine tuning or training neural language models, and more complicated feature engineering of samples, including line segmentation, contrast, and greyscale conversion. While utilities like HuggingFace have greatly simplified this process, the CNN's older and less

1 ' ' that this plan of the Duke of Portland's for crowding Gaols and taxing Counti es was not in every point of view a new one, appears from a Memorial presented by the Middlesex Justices to his Grace about 16 Months before, on the subject of the State Prisoners committed to Cold Bath Fields. The Memorial is printed at length i n the "Papers presented "to the House of Commons relative to his Majesty's Prison "in Cold Bath Fields". Ordered b to be printed 18, Decr 1880. pp 78, 79. "His Maje sty's Justices of the Peace for the County of Middlesex ... do ... humbly entrea t Your Grace to take into "consideration the difficulties under which they labour, which "are yet farther encreased by the addition of so great a "number of Prisoners of the above description" [State Pri- -soners] "to those whom the House of Correc tion was originally "intended to receive, and with which it is crowded to "a gre at degree of inconvenience "They also beg leave to observe, that a very great "additional expence has been incurred, on account of Pri- -soners "of the above des cription; which expenditure, as "the various Burthenes upon the County Rate, are al ready "very severely felt, they are deeply concerned to be under "the necessity of adding to the ordinary Charges of the Prison; "but this Expence the Magistrates confidently hoped will "be reimbursed to the County, on a fair statement of the "several Particulars". - From the so often mentioned posterior letter of his Grac e's, which I have had the honour to bring to light, information in no small degre e interesting may now be derived by these same Magistrates. In May 1798 they comp lain of the "crowding" of their Gaol with Prisoners not intended for it, and of inconvenience, assume (as was but natural) that the inconvenience was, as such, u nintentional on the part of the Duke, and, in mentioning it in that character, con cieve that the consideration thus submitted to that Noble person's will, in the c haracter of a motive, give birth in his Grace's mind to a determination to grant the relief they pray for. Little did they expect to see the day, & that so ear ly a one, in which by a document under 28"

1 ' ' That the plan of the Duke of Portland for dividing Great and Little Portland St reets in every point of view appears to be very proper and convenient.\n\nhis Grace did appear from a Memorial presented by the said Owners about 18 Months before, on the subject of the late Riots relative to Cold Bath Fields.\n\nThe Memorialists humbly crave to lay before your Majesty their apprehensions relative to His Majesty 's Royal Palace in Cold Bath Fields, ordered to be erected 19th February 1788.\n\nThat the said Fields are the Place for the butchery and consideration of His Maj esty's oxen, which are driven to the said Fields under which they labour, which as it is intended to receive, and with it is converted into a great degree of inconveni ence.\n\nThey also have the honor of observing that a very great additional expen se has been incurred, on account of the several necessary buildings which are absolutely necessary to adding to the ordinary charges of the Poor.\n\nThat the expense they are already under, and the necessity of adding to the said Buildings particula rly with regard to the Chapel, on a fair statement of the several grievances to the County, in a few days will be laid before your Majesty.\n\nFrom the often mentioned Rioters of this information in no small degree interests the said Owners, they cr ave the honor to bring to light, in Cold Bath Fields, ordered to be erected 19th Fe bruary 1788.\n\nThat the said Fields are the Place for the butchery and considerat ion of His Majesty 's oxen, which are driven to the said Fields under which they la bour, which as it is intended to receive, and with it is converted into a great deg ree of inconvenience.\n\nThey also have the honor of observing that a very great a dditional expense has been incurred, on account of the several necessary buildings which are absolutely necessary to adding to the ordinary charges of the Poor.\n\nT hat the expense they are already under, and the necessity of adding to the said Bu ildings particularly with regard to the Chapel, on a fair statement of the several g rievances to the County, in a few days will be laid before your Majesty.\n\nFrom t he often mentioned Rioters of this information in no small degree interests the sai

Figure 7: Hallucinations in Gemini

documented code was far more complex to replicate. For cultural institutions and non-technical researchers, employing Gemini or other cloud hosted pretrained LLMs may greatly simplify workflows and reduce technical barriers to entry.

For future work on this project, I plan to also evaluate the cross corpus effectiveness of TrOCR, especially on texts from different periods and languages outside the training data. This will ideally evaluate the transfer learning capabilities of the model and provide a more fair comparision to the zero shot example of Gemini. I also plan to experiment with prompting techniques or model settings to reduce the hallucination prevalence of Gemini and other LLMs.

Acknowledgments

References

- [1] T. Causer, K. Grint, A.-M. Sichani, M. Terras, 'making such bargain': Transcribe bentham and the quality and cost-effectiveness of crowdsourced transcription, *Digital Scholarship in the Humanities* 33 (2018) 467–487.
- [2] K. Dutta, P. Krishnan, M. Mathew, C. Jawahar, Improving cnn-rnn hybrid networks for handwriting recognition, in: 2018 16th international conference on frontiers in handwriting recognition (ICFHR), IEEE, 2018, pp. 80–85.
- [3] A. Baldominos, Y. Saez, P. Isasi, A survey of handwritten character recognition with mnist and emnist, *Applied Sciences* 9 (2019) 3169.
- [4] N. Babu, A. Soumya, Character recognition in historical handwritten documents—a survey, in: 2019 international conference on communication and signal processing (ICCP), IEEE, 2019, pp. 0299–0304.
- [5] H. J. Dolfing, Whole page recognition of historical handwriting, *arXiv preprint arXiv:2009.10634* (2020).

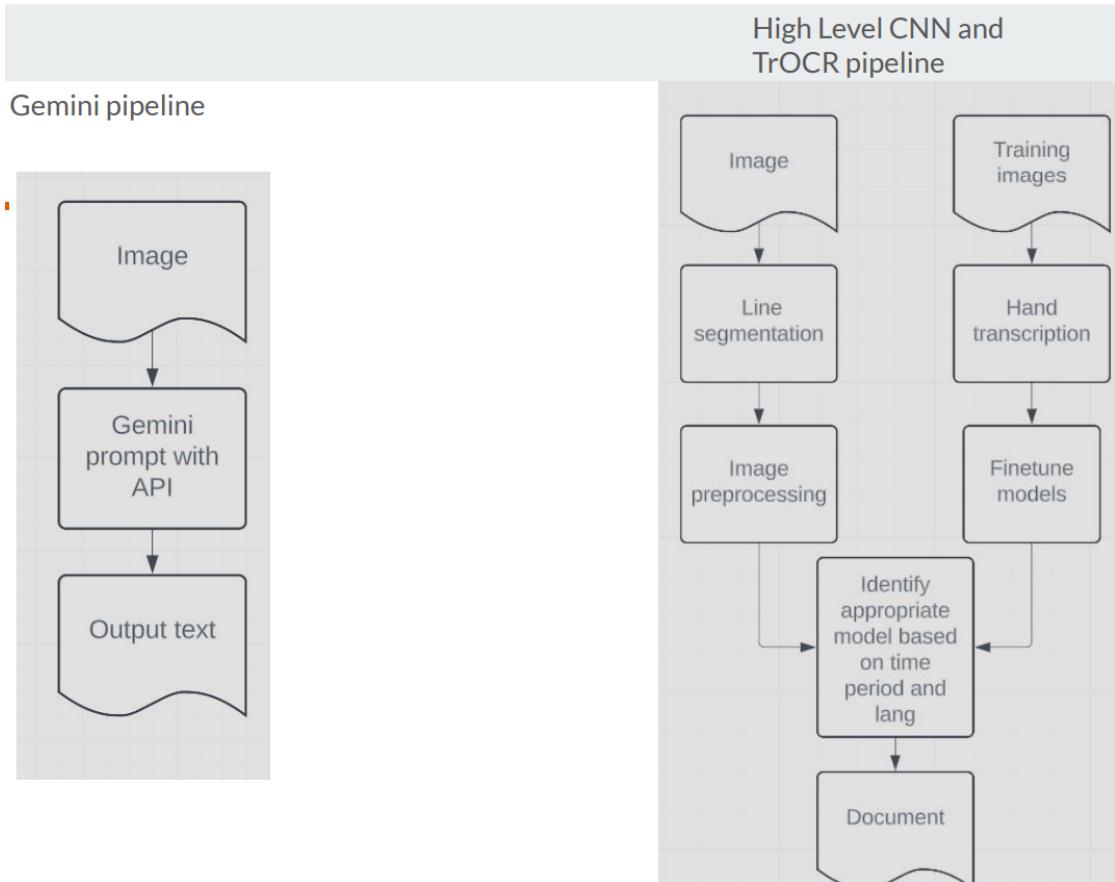


Figure 8: Gemini vs SoTA workflows

- [6] D. Parres, R. Paredes, Fine-tuning vision encoder–decoder transformers for handwriting text recognition on historical documents, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 253–268.
- [7] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, Trocr: Transformer-based optical character recognition with pre-trained models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.
- [8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
- [9] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, L. Wang, The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. arXiv: 2309.17421.
- [10] Y. Shi, D. Peng, W. Liao, Z. Lin, X. Chen, C. Liu, Y. Zhang, L. Jin, Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation, arXiv preprint arXiv:2310.16809 (2023).
- [11] J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, E. Vidal, A set of benchmarks for handwritten text recognition on historical documents, Pattern Recognition 94 (2019) 122–134.

- URL: <https://www.sciencedirect.com/science/article/pii/S0031320319302006>. doi:<https://doi.org/10.1016/j.patcog.2019.05.025>.
- [12] H. J. Dolfig, J. Bellegarda, J. Chorowski, R. Marxer, A. Laurent, The “scribblelens” dutch historical handwriting corpus, in: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2020, pp. 67–72.
 - [13] N. Serrano, F. Castro, A. Juan, The rodrigo database., in: LREC, 2010, pp. 19–21.
 - [14] E. Chammas, C. Mokbel, L. Likforman-Sulem, Handwriting recognition of historical documents with few labeled data, in: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), IEEE, 2018, pp. 43–48.

7. Online Resources

Github and datasets removed for anonymization purposes