Small-Scale Testing on Generative AI and Post-OCR Correction in Historical Datasets

Florentina Armaselu[1]

[1] Luxembourg Centre for Contemporary and Digital History (C^2DH), University of Luxembourg

1. Introduction

Recent developments in large language models (LLMs) and generative AI (GenAI) chatbots such as Chat-GPT, Google Bard (now Gemini), and YouChat (Brown et al., 2020; Chaka, 2023; Manyika and Hsiao, 2023) have fostered new types of interaction that can lower the barrier in human-machine communication through conversation in natural languages. We assume that such chatbots may be able to act as conversational assistants in tasks that otherwise require more complex processing to improve the results produced by simpler or earlier, less accurate techniques. This article proposes a set of small-scale tests with GenAI chatbots on post-OCR correction in historical datasets. It illustrates, through examples of responses obtained from GenAI agents integrated into post-OCR correction and assessment tasks, what types of challenges should be addressed in this context when working with historical datasets.

Previous studies have shown that OCR errors in input data can have a non-negligible impact on downstream language processing, such as sentence segmentation, named entity recognition (NER), topic modelling, word embedding, sentiment analysis and information retrieval (Nguyen et al., 2022; Strien et al., 2020). Therefore, various methods have been envisaged to tackle this problem, including post-OCR detection and correction of errors. Recent surveys classified these methods into manual and (semi-)automatic, isolated-word and context-dependent types, and emphasised the trend in the development of neural network- and context-based approaches (Nguyen et al., 2022). Other enquires focused on the use of machine learning techniques to automatically estimate text quality and select candidates for OCR rerun within cultural institutions that deal with lower quality historical data (Schneider and Maurer, 2022). On the other hand, studies on post-OCR error detection and correction investigated the use of large language models in this type of tasks. For example, pre-trained language models from the GPT-2 family were tested on datasets of English books and texts from post-OCR competitions by combining multiple OCR versions of an object and choosing the best-scored option, with the goal of reducing the number of errors (Gupta et al., 2021). Tests with other models, such as Llama 2, and a corpus of 19th century British newspapers containing aligned excerpts of machine and manual transcriptions, were also performed to assess the ability of prompt-based approaches to detect and correct OCR errors (Thomas et al., 2024). The performance of generative LLMs in post-OCR correction tasks was also examined by conducting experiments with GenAI agents such as ChatGPT-4, ChatGPT-4o, Claude 2.1, and Gemini 1.0 Pro and comparing their outputs when exposed to historical datasets in different languages and various types of prompt (Bang, 2024).

While these studies were generally based on the investigation of larger text collections, our set of tests explored the potential of using GenAI agents in post-OCR correction from a small-scale perspective. This involved the analysis of a use case and particular examples, with the aim of identifying potential categories of challenges and drawing a generic pipeline that may serve as a preliminary testbed for larger-scale applications. We assumed that early identification of challenges and prompt calibration scenarios from small-scale tests can eventually guide further developments. This type of enquiry was part of a larger project that used word embedding results (neighbour lists) and citations extracted from a selection of historical French monographs (1690-1918).[1] These extracts were integrated into a multilingual diachronic collection of interconnected terms expressed in RDF-XML, following a linguistic linked open data (LLOD) model (Armaselu et al., 2024).[2] In this context, the need for OCR correction of neighbour lists and citations has been identified. The article presents the applied methodology and its results in Section 2. Section 3 is dedicated to a discussion of the findings followed by a conclusion and remarks on future work in Section 4.

2. Methodology and results

The examples presented below were produced using three chatbots, ChatGPT-4, Google Bard, and YouChat, which were selected based on availability via subscription and free accounts and on reports considering them among the top AI chatbots (Davis, 2023). The conversations were intended to identify potential problems in the lists of neighbours of the word *révolution* extracted from the French monograph dataset by time slice, such as OCR errors (Table 1). The three chatbots were also asked to provide confidence levels for their corrections (in percentage).[3]

More complex prompts included several exchanges with the GenAI agent (ChatGPT-4) when asking for OCR error identification in citations extracted from the corpus corresponding to a given sense of the word *révolution* (Figure 1). Information about the origin of the excerpts, that is, French historical texts and their publication year or period, was also included in the prompts. Figure 1 (right) shows the OCR-ed text that was sent to ChatGPT-4 for correction.

Following this preliminary testing phase, a small-scale assessment of ChatGPT-4 was also performed using test and gold standard excerpts from the ICDAR 2017 competition on post-OCR text correction (Chiron et al., 2017; Rigaud et al., 2019) that included English and French monographs and periodicals from 1654 to 1934. The fragment contained 50 sentences from a text by Montesquieu, *Arsace et Isménie*, published in 1783, chosen to be close to the publication date of the text shown in Figure 1. Both the test fragment and the gold standard were extracted from the ICDAR2017 dataset of French monographs.[4]

1 The MONOGRAPH TEXT PACK from the BnL Open Data Collection. `https://data.bnl.lu/data/historical-newspapers/` (accessed 10 February 2025).

2 The LLODIA model is available at: `https://github.com/nexuslinguarum/LLODIA`.

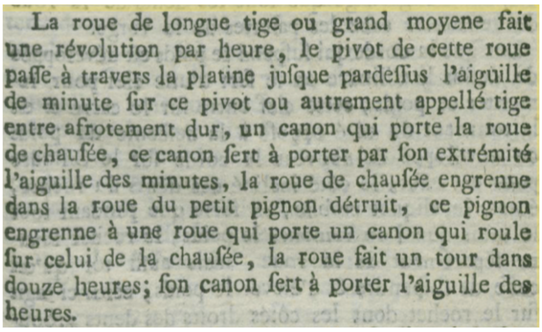
3 The chatbot confidence levels were considered as indicative, especially in cases when the agent changed its response in its interaction with the user. More research is needed to evaluate this type of chatbot self-assessment.

4 `https://sites.google.com/view/icdar2017-postcorrectionocr/dataset` (accessed 10

Table 1: Identifying OCR errors. Prompt: Please try to identify the words that may be due to OCR errors in the following list: *vandalisme, insurrection, insurgé, confération, lirnbourg, eévolution, vindépendance, holstein, usurper, exéquatur*.

Chatbot	**Interaction date**	**Answer excerpts**	**Chatbot confidence**
ChatGPT-4	15.12.2023	• confération - Possible Correction: "confédération" (confederation)	85%
		• lirnbourg - Possible Correction: "Limbourg" (Limburg)	80%
		• eévolution - Possible Correction: "révolution" (revolution)	95%
		• vindépendance - Possible Correction: "indépendance" (independence)	90%
Bard	18.12.2023	• confération - confédération	100%
		• lirnbourg - Luxembourg	95%
		• eévolution - évolution	90%
		• vindépendance - indépendance	80%
YouChat	16.12.2023	• "lirnbourg" to "Limbourg"	60%
		• "eévolution" to "évolution"	80%
		• "vindépendance" to "indépendance"	90%

ChatGPT-4 was asked to correct the first 50 sentences of the test fragment that was supposed to contain OCR errors. Then, it was asked to compare the initial and corrected texts with the corresponding gold standard (GS), to calculate the character error rate (CER), word error rate (WER) (Abadie et al., 2022), and the corresponding accuracy values.[5]


La roue de longue tige ou grand moyene fait une révolution par heure, le pivot de cette roue paſſe à travers la platine juſque pardeſſus l'aiguille de minute ſur ce pivot ou autrement appellé tige entre afrotement dur, un canon qui porte la roue de chauſée, ce canon ſert à porter par ſon extrémité l'aiguille des minutes, la roue de chauſée engrenne dans la roue du petit pignon détruit, ce pignon engrenne à une roue qui porte un canon qui roule ſur celui de la chauſée, la roue fait un tour dans douze heures; ſon canon ſert à porter l'aiguille des heures.

La roue de longue tige ou gand moyene fait une révolution par heure, le pivot de cette rous pale à travers la platune jufque pardeffus l'aiguille de minute 1ur ce pivot ou autrement appellé tige entre afrotement dur, un canon qu porte la roue de chaufée, ce canon fert à porter par fon extrémit l'aiguille des minuts, la roue de chaufée engrenne dans la roue du petit pignon détruit, ce pignon engrenne à uns roue qui porte un canon qui roule fur celui de la chaufée, la roue fäit un tour dans douze heures; fon canon fert à porer l'aiguille des neures.

Figure 1: Extract (Rosset, 1789, p. 13): a) image; b) OCR-ed text (French).

Table 2 shows two iterations of this process, since it was observed that ChatGPT-4 had a tendency to correct historical word forms in French, and additional prompting was necessary to prevent this type of change.[6] For comparison, the CER and WER

February 2025). The first 50 sentences were extracted from the full dataset, fr_monograph sub-folder, file 8.txt. Some preparations were necessary, for example to remove the symbol '@' from the excerpts, which was used in the competition for alignment at the character level. The sentences were also manually separated, one per line, and numbered from 1 to 50 to facilitate the interpretation of the results.

5 This was done by subtraction of the error rates from 100.

6 ChatGPT-4 was selected for this type of task since it allowed downloading the files corresponding to the various versions to be compared. Due to instability in results and reported errors by the GenAI agent, the calculation of the CER and WER values was repeated several times (on 31.01.2024, 07.02.2024,

Table 2: ChatGPT-4 CER, WER (with jiwer references and P, R, Fm) and examples (ICDAR 2017 excerpts).

Test excerpts, ICDAR and ChatGPT-4 corrected versions after iteration 1 and 2	**Interaction date**	**CER, P, R, Fm**	**WER, P, R, Fm**	**Chatbot confidence**	**Differences in the analysed text / GS**
OCR_toInput_ Arsace_et_ Isménie_1783 (ICDAR test fragment to be denoised, 983 tokens)	03.05.2024	1.22% *1.38%* (*jiwer*) P = 0.98 R = 0.99 Fm = 0.99	6.68% *7.55%* (*jiwer*) P = 0.92 R = 0.94 Fm = 0.93	95%	".on" / "son" "pres-que" / "presque" "Ambassa-deurs," / "Ambassadeurs,"
ChatGPT-4_ Corr1_Arsace_ et_Isménie_1783 (ICDAR test fragment after correction 1 by ChatGPT-4, 980 tokens)	03.05.2024	5.44% *5.15%* (*jiwer*) P = 0.95 R = 0.97 Fm = 0.96	18.22% *18.01%* (*jiwer*) P = 0.83 R = 0.84 Fm = 0.83	95%	"règne" / "regne" "d'Artamène, " / d'Artamene, " "laissant" / "laissa" "désirait" / "desiroit"
ChatGPT-4_ Corr2_Arsace_ et_Isménie_1783 (ICDAR test fragment after correction 2 by ChatGPT-4, 977 tokens)	03.05.2024	3.23% *3.00%* (*jiwer*) P = 0.97 R = 0.99 Fm = 0.98	12.63% *12.31%* (*jiwer*) P = 0.88 R = 0.89 Fm = 0.88	95%	"règne" / "regne" "d'Artamène," / "d'Artamene," "et" / "1" "âme" / "ame"

values (in italics) were computed independently by a program that we have written for this purpose and that used the Python library jiwer.[7]

For each of the three analysed texts, we used the sentence alignments and number of hits or correct items (H), substitutions (S), deletions (D) and insertions (I) provided as output by the jiwer-based program. With these values, we computed more general evaluation metrics, precision (P), recall (R), and F-measure (Fm), corresponding to the two types of error calculated at the character and word level. The formulae for calculating P, R and Fm were inspired by general approach descriptions (Abadie et al., 2022; Karpinski et al., 2018; Morris et al., 2004) and adapted to our task and data as detailed below, [8]

$$P = H/(H + S + I) \quad (1)$$
$$R = H/(H + S + D) \quad (2)$$
$$Fm = 2PR/(P + R) \quad (3)$$

where P is the number of hits divided by the total items in the hypothesis (text to

02.05.2024 and 03.05.2024). The last results, which seemed more stable and closer to the jiwer-calculated values, were included in the table.

[7] https://jitsi.github.io/jiwer/ (accessed 10 February 2025).

[8] The values presented in Table 2 were computed in Excel using the information collected manually from the output files of the Python program, and formulae 1-3.

be assessed), R is the number of hits divided by the total items in the reference (GS), and Fm is the harmonic mean of P and R.

3. Discussion

A qualitative analysis of the chatbot's responses for the first task (Table 1) indicates that the three AI agents were quite successful in identifying possible errors. The word similarly identified by all the chatbots was "vindépendance", while slight differences in agreement can be observed for the other words. ChatGPT-4 agreed with Bard on "confération" and with YouChat on "lirnbourg", while Bard and YouChat agreed on "eévolution". Additional prompting was needed to remind Bard that corrected words should be provided in French (A: Table 4), since the proposed corrections were initially expressed in English. This behaviour was not observed for the other two agents that provided the answers directly in French (YouChat) or in French with English translations (ChatGPT-4).

When the prompt included more details about specific types of error, such as the use of elongated 's' in older texts, often confused by the OCR software with the letter 'f', the chatbots were able to propose plausible corrections. For example, YouChat after first proposing 'fantastic' as a correction of 'fuppofant', re-examined it in favour of the form "supposant" (confidence 70%) (A: Table 5) and re-assessed the level assigned to the initial choice "fantastic" (confidence 40%). ChatGPT-4 identified the correct form "supposant" (confidence 95%) and explained that in this case the long 's' was probably misinterpreted as 'f' in the OCR process.[9]

More complex prompts included several exchanges when asking for error identification in citations extracted from the corpus corresponding to a given sense of the word révolution (Figure 1). Although errors such as "platune" instead of "platine", "jufque pardeffus" instead of "jusque pardessus", or "fert" instead of "sert" were easily identified (confidence 90%), the form "chaufée" was more difficult to analyse and required additional prompting. After asking ChatGPT-4 to reconsider the form to take into account the elongated 's' issue, the context "roue de chaufée" and the possibility that it may refer to clock making in the 18th century, the chatbot proposed the correction "roue de chaussée", with "'chaussée' potentially referring to a part or a process in the clock's mechanism" (confidence 75%) (A: Table 6). Other difficulties consisted in preventing corrections of possible historical word usages, which were not due to OCR errors, such as "moyene", "engrenne", "afrotement" and their transformation into modern forms (A: Table 7). It can be noted that in response to prompts related to the historical use of certain words, the agent proposed plausible corrections, accompanied by explanations of the most likely modern forms, their English translations, and potential contextualisation from a historical point of view.

The same issue related to historical forms was observed when comparing the corrected ChatGPT-4 version with the ICDAR 2017 gold standard (Table 2). The lowest error and highest values of P, R, and Fm correspond to the ICDAR test fragment. For the two iterations using ChatGPT-4, the error rate decreased and the P, R, Fm increased for correction 2 as compared with correction 1 after the use of specific prompts indicating that historical forms should be preserved.[10] Regarding ICDAR 2017 results, (Chiron et al., 2017) reported that the best performing method of the competition in

---

9 Google Bard was not tested for this particular form.

10 Example of such prompt: *Please try to correct only the clear OCR errors, such as misrecognised characters and incorrect word splits, while retaining the historical French forms, such as "reposoit", "faudroit".*

terms of F-measure (Fm 0.55, P 0.51, R 0.59) for historical French monographs was WFST-PostOCR based on weighted finite-state edit transducers and bigram language models derived from the Google Books Ngram Corpus. The values shown in Table 2 for these measures and the texts we analysed are much higher. However, it should be considered that the sample used in the French monograph competition counted 32274 tokens with an average character error rate of 5%, higher than the text size and initial error rate from our small-scale test. These evaluation measures were calculated for task 1 (error detection), in which the competition participants were asked to provide only the position and length of the alleged erroneous tokens in the raw OCR-ed text, which differs from our full correction task.[11] Therefore, these values are not directly comparable with our figures, but may be used as indicative reference points. It should also be noted that some inaccuracies in the GS itself were reported by ICDAR participants, according to the article cited. Such inaccuracies can be observed as well in Table 2 (column 6) that shows some diacritics missing from the GS version while they are properly rendered in the ChatGPT-4 corrections.

An additional test was performed to compare the texts corrected by ChatGPT-4 with a more modern version (1876) (MD) of the fragment from *Arsace et Isménie*.[12] The calculated CER and WER values using jiwer for the GPT-corrected texts using this text as a reference[13] were higher compared to the values based on the gold standard in Table 2. Thus, MD-based rates compared to GS-based rates computed via jiwer showed an increase of 0.37% in CER and 2.35% in WER for correction 1, and respectively 2.41% in CER and 10.56% in WER for correction 2. This was surprising since given the agent's tendency to modernize the word forms, a closer resemblance with the MD version and lower error rates based on the comparison with it were expected. Some mismatches between the GenAI corrections and the reference versions were due to the use of different types of punctuation marks (e.g., straight vs. curly apostrophes, colon, semicolon and comma used differently) and capitalisation (e.g., Reine vs. reine). A closer look at the differences also showed that the ChatGPT-4 corrections not only targeted presumed OCR errors, but sometimes involved changing the order of words or rephrasing.[14] These findings align with recent studies reporting that GenAI agents, such as ChatGPT-4, ChatGPT-4o, Claude 2.1, and Gemini 1.0 Pro, show a certain tendency towards oversimplification or modernisation when applied to post-OCR correction of historical texts, which may impact the degree of accuracy and fidelity of the corrections as compared with the original content (Bang, 2024, pp. 40, 46).

Table 3 summarises the categories of differences observed between the versions

[11] ICDAR 2017 also included task 2 (correction of OCR errors) in which the participants were asked to provide a ranked list of replacement candidates for each error detected during task 1. The official metric for this second task consisted of the improvement rates with respect to the weighted sum of the Levenshtein distances between the correction candidates and the corresponding token in the GS, considered in fully- and semi-automated scenarios (Chiron et al., 2017).

[12] `https://fr.wikisource.org/wiki/Arsace_et_Ism%C3%A9nie` (accessed 10 February 2025).

[13] MD reference, ChatGPT-4 correction 1 (5.52% CER and 20.36% WER); correction 2 (5.41% CER and 22.87% WER).

[14] For instance, "Il écrivait à la Reine les lettres les plus tendres du monde" (corr. 1) vs. "Il écrivoit à la Reine les lettres du monde les plus tendres" (GS, corr. 2) vs. "Il écrivait à la reine les lettres du monde les plus tendres" (MD); "Elle descendit de son char et entra dans le temple" (corr. 1) vs. "Elle descendit de son char, entra dans le temple" (GS, corr. 2) vs. "Elle descendit de son char, et entra dans le temple" (MD); "il cherchait l'étranger et le trouva plongé dans une affreuse tristesse" (corr. 1) vs. "il cherchoit l'étranger, il le trouva dans une affreuse tristesse" (GS) vs. "il cherchoit l'étranger, et le trouva dans une affreuse tristesse" (corr. 2) vs. "il cherchait l'étranger, et il le trouva dans une affreuse tristesse" (MD).

Table 3: Diff examples classified by ChatGPT-4o and its confidence levels (07.10.2024).

Diff category	**Examples (GS / ChatGPT-4_Corr1)**	**Chatbot confidence**
Punctuation	*Soyez mon ami*, lui dit-il, *puisque vous êtes malheureux.* / "Soyez mon ami", lui dit-il, "puisque vous êtes malheureux."	95%
Orthography	Cette entrée fut singuliere. / Cette entrée fut singulière.	85%
Paraphrasing	Ce Prince mourut accablé d'ennuis, laissa** son trône à sa fille Isménie. / Ce Prince mourut accablé d'ennuis, laissant son trône à sa fille Isménie.	85%
Modernisation, orthography, syntax/structure	Il menoit avec lui tout ce qui étoit propre à un appareil de nôces **des danseurs, des joueurs d'instruments, des farceurs, des cuisiniers, des eunuques, des femmes il menoit avec lui une formidable armée. / Il menait avec lui tout ce qui était propre à un appareil de noces : des danseurs, des joueurs d'instruments, des farceurs, des cuisiniers, des eunuques, des femmes, ** *e***t********* une formidable armée.	90%

corrected by ChatGPT-4 and the two references (GS and MD).[15] To compile the table, we used the output files of the Python program for CER and WER computing, which contained character-level alignments for pairs of aligned sentences between the compared versions, and difference markers (for substitution, insertion, deletion).

We prompted ChatGPT-4o, with the diff files attached to the prompts, to assist in classifying the sentence pairs according to six categories of differences. As illustrated in the table, the analysed sentences can belong to single or multiple categories. The diff categories were defined as follows: *spacing* (differences in missing or extra-spaces between characters or words), *punctuation* (use of different punctuation marks), *orthography* (differences at the character level, such as diacritics or minor spelling variations in a word form), *modernisation* (replacement of older word forms by modern equivalents), paraphrasing (use of different words or phrases intended to convey a similar meaning), *syntax/structure* (changes in sentence structure through word insertion, deletion or reordering). This type of analysis can draw paths for further enquiry regarding the use of GenAI technology in two different types of tasks, post-OCR correction and modernisation of historical texts. While some changes can be imagined as potential ingredients of a modernisation pipeline, others should be prevented by explicit specifications included in the prompts to preserve the original forms as much as possible.

Based on these observations, we argue that GenAI-based workflows for post-OCR correction should include preliminary small-scale testing and prompt calibration to enable early identification of accuracy issues and their potential solutions. Figure 2

[15] The experiment involved several iterations with ChatGPT-4o for the refinement of the categories and classification of sentence pairs according to these categories. Four files were attached to the prompts to be analysed. They contained aligned sentences at the character level for the ChatGPT-4_Corr1 and ChatGPT-4_Corr2 as compared with the GS and MD versions. Some simplifications were applied to the categories proposed by ChatGPT-4o to avoid category overlapping. The character '*' was used by the Python program for alignment, in case of differences between sentences.

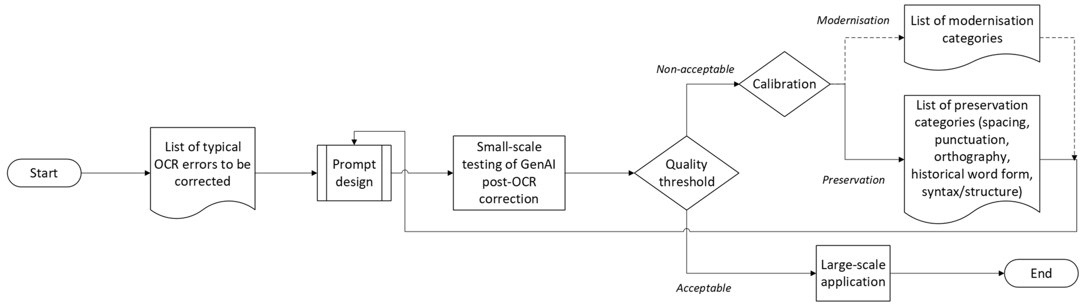


Figure 2: Generic pipeline for small-scale testing of GenAI post-OCR correction.

illustrates a generic pipeline that considers both the preservation and modernisation branches, taking into account the various types of OCR errors due to misinterpretation of letters, specific or older fonts, print quality, etc., and the categories of differences between the original and corrected text resulting from the particularities of the generative AI technology.

One of the typical emergent abilities of large language models, which are not present in smaller pre-trained models, is in-context learning. Compared to fine-tuning that implies adaptation of a model to a specific task by training it on a tailored dataset and updating its weights, or to retrieval-augmented generation (RAG) that combines information retrieval and domain-relevant knowledge bases, in-context learning involves the use of zero-, one- or few-shot learning including in the prompts task descriptions with or without examples, and no update of the model parameters (Brown et al., 2020; Li, 2023; Zhao et al., 2024). While various fine-tuning and RAG techniques are available (Parthasarathy et al., 2024), in-context learning may represent an approach that is potentially interesting to explore for tasks such as LLM-based post-OCR correction. Recent multilingual benchmarks have shown both promise and challenges in applying LLMs in zero- and few-shot settings to post-OCR correction of historical datasets (Boros et al., 2024; Kanerva et al., 2025), some of them similar to observations presented in this paper (e.g., tendency of the models to overcorrect and paraphrase). The decision on which of these categories of adaptation techniques to choose also depends on the complexity of the project and the budget allocated to it. Although the utility of large-scale benchmarking is undeniable, small-scale testing, including data preparation, processing, and evaluation, and iterative prompt tuning, can prove its usefulness as a faster and cheaper way of identifying categories of problems from the outset, suggesting directions for adaptation, further investigation, or assessment aligned with existing standards, that may inform the design of subsequent larger-scale phases or applications.

4. Conclusion and Future Work

The article proposes a small-scale investigation on the use of GenAI agents in post-OCR correction workflows for historical datasets. Although preliminary results show a certain potential for this type of technology in solving tasks in this category, more tests are necessary to assess the capacity to respond to specially conceived prompts for historical text processing. In particular, it was shown that the agents have a tendency to replace historical forms with more modern ones, to reformulate whole phrases, or to change word order and punctuation. Specific instructions should be devised to prevent these forms of modification. However, building modernised layers for documents with old spelling may be considered a possibly interesting application in tasks such as the transformation of older texts to be read by modern users or computer programs.

Another potential challenge resides in a certain degree of instability observed in computing character and word error rates (CER, WER). This indicates that comparing the GenAI values with independently computed results using programming languages such as Python or R should be envisaged. The study was limited to interactions with three agents that involved the use of online platforms, while the integration of this type of technology into larger-scale pipelines would require more code-oriented or batch processing solutions that need to be further examined. Future work can include additional testing with other GenAI agents, open-source large language models in a programming environment, datasets in other languages, and further enquiry about the application of these technologies in post-OCR correction for historical and digital humanities research. Although limited in scope, it may be argued that small-scale testing of the kind presented in this paper can serve as a pilot phase in the pipeline, that allows for early identification of problems, synthesis of observations, assessment, and articulation of hypotheses for further testing or potential solutions, which can inform the design of subsequent larger-scale phases in the workflow or applications.

References

N. Abadie, E. Carlinet, J. Chazalon, and B. Duménieu. *A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories*, volume 13237 of *Lecture Notes in Computer Science*, page 445–460. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06554-5. doi: 10.1007/978-3-031-06555-2_30. URL `https://link.springer.com/10.1007/978-3-031-06555-2_30`.

Florentina Armaselu, Barbara McGillivray, Chaya Liebeskind, Paola Marongiu, Giedrė Valūnaitė Oleškevičienė, Elena-Simona Apostol, and Ciprian-Octavian Truică. Multilingual word embedding and linguistic linked open data for tracing semantic change. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 50(2):219–257, 2024. ISSN 18490379, 13316745. doi: 10.31724/rihjj.50.2.1.

Kyunga Bang. *Exploring Generative Large Language Models for Post-OCR Enhancement of Historical Texts*. Master's thesis, Graduate School of Sungkyunkwan University, April 2024. URL `https://www.researchgate.net/publication/382113790_Exploring_Generative_Large_Language_Models_for_Post-OCR_Enhancement_of_Historical_Texts`.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of LaTeCH-CLfL 2024*, page 133–159. Association for Computational Linguistics, March 2024. URL `https://aclanthology.org/2024.latechclfl-1.14/`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. (arXiv:2005.14165), July 2020. URL `http://arxiv.org/abs/2005.14165`. arXiv:2005.14165 [cs].

Chaka Chaka. Generative AI chatbots - ChatGPT versus YouChat versus Chatsonic: Use cases of selected areas of applied English language studies. *International Journal of Learning, Teaching and Educational Research*, 22(6):1–19, June 2023. ISSN 16942493, 16942116. doi: 10.26803/ijlter.22.6.1.

Guillaume Chiron, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. ICDAR2017 competition on post-OCR text correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, page 1423–1428, Kyoto, November 2017. IEEE. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.232. URL `http://ieeexplore.ieee.org/document/8270163/`.

Krissy Davis. The best AI chatbots: ChatGPT and other alternatives, May 2023. URL `https://www.wearedevelopers.com/en/magazine/236/best-ai-chatbots-chatgpt-and-other-alternatives`.

Harsh Gupta, Luciano Del Corro, Samuel Broscheit, Johannes Hoffart, and Eliot Brenner. Unsupervised multi-view post-OCR error correction with language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 8647–8652, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.680. URL `https://aclanthology.org/2021.emnlp-main.680`.

Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho, and Filip Ginter. OCR error post-correction with LLMs in historical documents: No free lunches. (arXiv:2502.01205), February 2025. doi: 10.48550/arXiv.2502.01205. URL `http://arxiv.org/abs/2502.01205`. arXiv:2502.01205 [cs].

Romain Karpinski, Devashish Lohani, and Abdel Belaid. Metrics for complete evaluation of OCR performance. In *The 22nd Int'l Conf on Image Processing, Computer Vision, & Pattern Recognition*, Las Vegas, United States, July 2018. URL `https://inria.hal.science/hal-01981731v1`.

Yinheng Li. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, page 641–647, 2023. doi: 10.26615/978-954-452-092-2_069. URL `http://arxiv.org/abs/2309.13205`. arXiv:2309.13205 [cs].

James Manyika and Sissie Hsiao. *An overview of Bard: an early experiment with generative AI*. 2023. URL `https://ai.google/static/documents/google-about-bard.pdf`.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, page 2765–2768. ISCA, October 2004. doi: 10.21437/Interspeech.2004-668. URL `https://www.isca-archive.org/interspeech_2004/morris04_interspeech.html`.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. Survey of post-OCR processing approaches. *ACM Computing Surveys*, 54(6):1–37, July 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3453476.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. (arXiv:2408.13296), October 2024. doi: 10.48550/arXiv.2408.13296. URL `http://arxiv.org/abs/2408.13296`. arXiv:2408.13296 [cs].

Christophe Rigaud, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. ICDAR 2019 competition on post-OCR text correction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, page 1588–1593, Sydney, Australia, September 2019. IEEE. ISBN 978-1-7281-3014-9. doi: 10.1109/ICDAR.2019.00255. URL `https://ieeexplore.ieee.org/document/8978127/`.

F. Rosset. *L'art de conduire et regler les pendules et les montres*. chez la Veuve de J. B. Kleber, Imprimeur de Sa Majesté, Luxembourg, 1789. URL `https://viewer.eluxemburgensia.lu/ark:70795/dqgfr3/pages/17/articles/DTL612`.

Pit Schneider and Yves Maurer. Rerunning OCR: A machine learning approach to quality assessment and enhancement prediction. *Journal of Data Mining & Digital Humanities*, November 2022. ISSN 2416-5999. doi: 10.46298/jdmdh.8561. URL `http://arxiv.org/abs/2110.01661`. arXiv:2110.01661 [cs].

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, page 484–496, Valletta, Malta, 2020. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-395-7. doi: 10.5220/0009169004840496. URL `http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496`.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. Leveraging LLMs for post-OCR correction of historical newspapers. In *LT4HALA 2024@LREC-COLING-2024*, page 116–121. ELRA Language Resources Association, 2024. URL `https://aclanthology.org/2024.lt4hala-1.14.pdf`.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. (arXiv:2303.18223), October 2024. doi: 10.48550/arXiv.2303.18223. URL `http://arxiv.org/abs/2303.18223`. arXiv:2303.18223 [cs].

A. Appendix

Table 4: Prompt and response excerpts, reminding the language (Google Bard, 18.12.2023).

User	GenAI agent (Bard)
Please try to identify the words that may be due to OCR errors in the following list and assign a confidence level in percentage to each proposed correction. The words are in French and the corrections should also be in French. …	Sure, here are the words in the list that may be due to OCR errors, along with proposed corrections and confidence levels: … confération confédération 100% …

Table 5: Prompt and response excerpts, “supposant” (YouChat, 16.12.2023).

User	GenAI agent (YouChat)
Would your decision be different if you consider that usually elongated ’s’ in older texts is producing OCR errors, such as considering ’f’ instead of ’s’?	Considering the possibility of an OCR error substituting "f" for elongated "s" is indeed a relevant factor to consider. Given this information, the potential correct form for the word "fuppofant" might indeed be "supposant" if we consider the substitution of "s" with "f" due to OCR errors in older texts. Taking this into account, the corrected potential form for "fuppofant" could be "supposant" instead of "fantastic." …

Table 6: Prompt and response excerpts, "roue de chaussée" (ChatGPT-4, 15.12.2023).

User	GenAI agent (ChatGPT-4)
Can you please reanalyze the word "chaufée" given the context "roue de chaufée" and the fact that the text refers to clock making in the 18th century?	In the context of clockmaking during the 18th century, the term "chaufée" in "roue de chaufée" could be a historical or technical term related to the mechanism of a clock. Considering the OCR misinterpretation of long ’s’ as ’f’, and given the clockmaking context, the term could be "roue de chaussée", with "chaussée" potentially referring to a part or a process in the clock’s mechanism. However, without specific knowledge of 18th-century clockmaking terminology, it’s difficult to ascertain the exact term. "Roue de chauffée" could refer to a heated part or a wheel that is treated in a particular way, but this seems less likely in the context of mechanical clockwork where terms like "chaussée" do not clearly fit. Considering this, a more likely correction would be "roue de chaussée" if "chaussée" is a term used in the context of clock mechanics at the time. However, if "chauffée" is not due to an OCR error and is actually a correct transcription of the original text, it might require specialized historical knowledge to interpret accurately. Based on common French usage and the potential for an OCR error, "roue de chaussée" could be a plausible correction with a medium to high confidence level, but without specific expertise in historical clockmaking terminology, I cannot provide a definitive correction.

Table 7: Prompt and response excerpts, "afrotement", "engrenne" (ChatGPT-4, 15.12.2023)

User	**GenAI agent (ChatGPT-4)**
Do you have any information about the word "afrotement ", in the text "tige entre afrotement dur", and its possible usage in French in the 18th century? Same question for "engrenne ", in "ce pignon engrenne". It is possible that these are historical usages that should be kept in the transcription of the text. The goal is to correct just the errors that may be due to OCR misinterpretation.	In the 18th century, French language and spelling were less standardized than they are today, and many words could have various spellings or be spelled according to phonetics. With this in mind: • Afrotement: The correct modern French word is "frottement," which means "friction." It is plausible that "afrotement" is a historical spelling or a typographical error in the original text. If "afrotement" is not an OCR misinterpretation, it could be an archaic spelling of "frottement" that was used at the time. This would be especially true in the context of clock mechanisms where friction is an important concept. • Engrenne: The modern French word "engrène" comes from the verb "engrener," which means "to mesh" or "to engage," as gears do in a clock. "Engrenne" could be an old conjugation or spelling of the verb. If "engrenne" is correctly transcribed, it may not be an OCR error but rather a historical usage that should be preserved in the text. …