Post-OCR Correction of Historical German Periodicals using LLMs

Vera Danilova and **Gijs Aangenendt**
Uppsala University
Dept. of History of Science and Ideas
Thunbergsvägen 3P, 752 38, Uppsala, Sweden
first_name.last_name@idehist.uu.se

Abstract

Optical Character Recognition (OCR) is critical for accurate access to historical corpora, providing a foundation for processing pipelines and reliable interpretation of historical texts. Despite advances, the quality of OCR in historical documents remains limited, often requiring post-OCR correction to address residual errors. Building on recent progress with instruction-tuned Llama 2 models applied to English historical newspapers, we examine the potential of German Llama 2 and Mistral models for post-OCR correction of German medical historical periodicals. We perform instruction tuning using two configurations of training data, augmenting our small annotated dataset with two German datasets from the same time period. The results demonstrate that German Mistral enhances the raw OCR output, achieving a lower average word error rate (WER). However, the average character error rate (CER) either decreases or remains unchanged across all models considered. We perform an analysis of performance within the error groups and provide an interpretation of the results. The code and resources are publicly available.[1]

1 Introduction

The effectiveness of transcription methods, such as optical character recognition (OCR), in processing historical documents critically influences the accuracy of search and analysis in text processing pipelines (Lyu et al., 2021). Despite advances in OCR technology, library and archive collection transcriptions often contain significant errors and noise due to factors such as scan quality, language, layout complexity, and character similarity. Inaccuracies in OCR transcriptions can propagate through multistep historical text processing pipelines, hinder performance on downstream Natural Language Processing (NLP) tasks, and create a risk of distorted interpretations (Lopresti, 2008; van Strien et al., 2020). Post-OCR correction plays an important role in mitigating these errors and improving transcription quality.

We focus on the post-correction of a dataset from an ongoing project[2] on the modern history of medicine, which explores ten European patient organizations. In this paper, we consider the periodical of the German Diabetes Association "Der Diabetiker", issued between 1951 and 1990. The materials predate the German spelling and punctuation reform of 1996, when new rules were implemented regarding the double s (ß), consonants, capitalization, hyphenation, and loanwords, making the dataset different from modern texts. The quality of raw OCR output varies significantly, with simpler layouts achieving higher accuracy, while complex multicolumn layouts containing advertisements and rare fonts often result in numerous errors.

In this paper, we address the following research questions:

1. Can the previously successful approach for post-OCR correction of an English-language historical newspaper dataset (Thomas et al., 2024) be effectively adapted using German-specific models? Additionally, will generative models outperform BART (Lewis et al., 2020) in reducing key metrics like the average Character Error Rate (CER) and Word Error Rate (WER)?

2. How does augmentation with a different source (National Library dataset including re-

[1] https://github.com/veraDanilova/ocr_post-correction_RESOURCEFUL-2025

[2] http://actdisease.org

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 120–129

ligious and cultural articles) contribute to the quality of post-OCR correction?

3. Given that our dataset includes both challenging pages with high initial CER and easier pages with near-perfect recognition, can post-correction improve difficult errors without compromising the quality of already well-recognized pages?

This paper unfolds as follows. Section 2 describes prior research. In Section 3, we present our annotated dataset alongside the augmentation datasets. Section 4 lays out the experimental setup. Finally, Section 5 discusses our findings and Section 6 concludes the paper.

2 Related Work

Post-OCR correction of historical documents has become a central theme at the International Conference on Document Analysis and Recognition (ICDAR). The conference hosted two competitions in 2017 and 2019 dedicated to post-OCR correction, introducing two key tasks: error detection and error correction. Sequence-to-sequence neural machine translation emerged as the dominant methodology among the most successful approaches showcased at this conference (Chiron et al., 2017; Rigaud et al., 2019). The authors of the competition emphasize that historical newspapers and periodicals continue to pose a substantial challenge to OCR systems, mainly due to their intricate layouts and typographic diversity (Rigaud et al., 2019). Following the conclusion of these competitions, the benchmarks were further utilized to advance the state of the art in post-OCR correction of newspapers with pre-trained models, specifically by finetuning BART (Soper et al., 2021).

Thomas et al. (2024) are the first to explore the instruction tuning of generative models for post-OCR correction of an English dataset of 19th century newspapers. Llama 2 models (Touvron et al., 2023) are reported to considerably outperform BART. The authors emphasize the adaptability of models like Llama 2 to downstream tasks with limited instruction-tuning data (Zhou et al., 2024) in contrast to machine translation models like BART that typically depend on large volumes of parallel data for optimal performance (Xu et al., 2024).

This paper addresses a real-world scenario involving a very limited annotated dataset of German historical medical periodicals, characterized by varying quality in the initial OCR. The dataset includes layouts and fonts that are easily recognized by models, as well as more complex layouts with distorted reading order, images, and advertisements featuring rare fonts and skewed text. Given the small size of this dataset, which precludes instruction tuning, we augment it with a German dataset from the ICDAR 2019 competition, which includes a similar time period and source - newspapers. Additionally, we explore augmentation using another ICDAR 2019 dataset, which represents a different source - cultural and religious materials from the German National Library.

This study does not explore augmentation with synthetic data. While artificially inserted errors can enhance model performance, they may fail to capture the complexity and diversity of real-world OCR errors, limiting the models' generalization ability (Jasonarson et al., 2023). This is particularly relevant for our dataset, where typical error insertion is insufficient due to the intricate challenges posed by complex layouts, such as those with advertisements. We leave the exploration of error generation approaches for our specific context to future work.

Our experiments contribute to post-OCR correction for German historical documents by comparing the performance of a finetuned German BART model with instruction-tuned German generative models, such as Llama 2 13b and Mistral 7b(Jiang et al., 2023). Beyond evaluating average performance metrics, we focus on error categories to better understand how the models handle specific types of errors and whether they degrade the quality in areas where OCR is already accurate.

3 Data

3.1 Der Diabetiker

The dataset contains pages from the patient organization periodical, Der Diabetiker (1951-1990), published by the German Diabetes Association[3]. The journal was digitized using ABBYY FineReader 14[4]. Deskew and straighten lines were

[3]The periodical changed name in 1971 to Diabetes-Journal

[4]https://www.abbyy.com/company/news/abbyy-finereader-14-pdf-solution/

selected as image processing steps in the workflow.

To create the ground truth, we manually corrected a sample consisting of 35 pages selected to represent layout complexity and time period. The quality of simple layouts is generally high, while most issues are concentrated in the more complex layouts. Pages considered as simple layout have only one or two columns, text in a common font (Times New Roman or Arial), and no advertisements or titles breaking the columns. Pages considered as complex layout contain full page advertisements, multi-text columns interspersed with advertisements and images, and rare fonts.

Overall, we collected 20 pages with complex layouts (12 pages from the period 1951-1970 and 8 pages from the period 1970-1990), and 15 pages with simple layouts (7 pages from the period 1951-1970 and 8 pages from the period 1970-1990).

3.2 Augmentation Datasets

To augment the training dataset, we utilize two ICDAR-19 competition datasets with ground truth for OCR post-correction: the Neue Zürcher Zeitung (NZZ) and the IMPACT German National Library dataset (GNL) [5].

The NZZ dataset includes 96 front pages of the Swiss newspaper Neue Zürcher Zeitung, covering the period from 1780 to 1947. Front pages were chosen because they typically contain highly relevant material. They include but not exclusively consist of advertisements.

The GNL dataset is a subset of the IMPACT dataset (Papadopoulos et al., 2013) that consists of 150 pages from various time periods. According to our manual analysis, it is mostly written in contemporary German, spanning different domains such as art, literature, and religion, with some excerpts in Latin. Neither the ICDAR-2019 competition nor the official description of the full version in Papadopoulos et al. (2013) provide detailed information on the distribution of time periods and domains within the German segment. However, the latter reports that the full version of the IMPACT dataset is predominantly composed of 19th-century data, accounting for 316k of the total 602k pages, followed by 20th-century data with 160k pages. More than half of the dataset consists of book pages (335k pages).

[5] https://zenodo.org/records/3515403

For NZZ and GNL datasets, special alignment files are provided to match OCR-ed text with ground-truth spans. Manual review of the aligned spans showed that in four NZZ pages and eleven GNL pages, the reading order was restored in the ground truth. Therefore, the OCR and ground truth spans are either partially or completely misaligned. Additionally, multiple pages exhibit partial mismatches due to missing text in the OCR output. The next section outlines the dataset types used to evaluate the impact of these misalignments.

4 Experimental Setup

In this study, we evaluate German BART and generative models, Llama 2 13b and Mistral 7b, comparing their WER and CER metrics[6] against raw OCR outputs.

At the core of the training process lies a base dataset consisting of Der Diabetiker pages, a small annotated collection, combined with NZZ, a comparable newspaper source. To evaluate the impact of dataset composition, we train the models with and without augmentation using GNL, which adds greater diversity to the data.

The training data is structured in two configurations: one that retains misaligned spans and another that excludes them.

To deepen our understanding of models' performance, we analyze their handling of diverse OCR errors across three distinct error categories.

Our primary focus is the correction of errors in the Der Diabetiker test data. The approach identified as successful will be further refined and expanded for application in post-OCR correction of the entire German segment of our project's dataset of patient organizations' periodicals.

4.1 Data Pre-processing

Pre-processing for all datasets includes removing extra spaces and duplicates. Additionally, we control for input context length based on the insights from previous work. For Der Diabetiker, we use the segmentation into paragraphs provided by the raw OCR output. For NZZ and GNL, the splitting strategy is detailed below.

[6] WER is the ratio of the minimum number of word substitutions, deletions, and insertions (word edit distance) required to transform the recognized text into the ground truth, divided by the total number of words in the ground truth. Similarly, CER is the character edit distance divided by the total number of characters in the ground truth.

Context length. We divided the NZZ and GNL pages into spans at newline characters, resulting in an average span length of 168 characters with a standard deviation of 32. This decision was motivated by prior work, which discussed the impact of text length on OCR post-correction (Veninga, 2024). Models like finetuned ByT5 (Xue et al., 2022) and Llama-2 7b, in zero-shot and few-shot settings, were found to be sensitive to context length. Long or very short spans make it challenging for these models to learn effectively.

To further investigate this, we analyzed results from prior work (Thomas et al., 2024) regarding OCR text length and CER reduction for the Llama 2 13b model. The table summarizing the results is provided by the authors in the associated GitHub repository[7]. It revealed that OCR texts exceeding 400 characters, though constituting a small fraction of the test set (38 out of 2792 texts), suffered a significant increase in errors (CER reduction = -190). At the same time, shorter spans showed notable improvement (CER reduction = 60). Given that the corresponding training set had an average text length of 124 characters, we decided to finetune on spans between 100 and 200 characters.

Training dataset configurations. To evaluate the impact of misalignments discussed in the previous Section on models' performance, we use two configurations of training sets for each of the datasets. ALL-DATA includes the full dataset without filtering, while FILTERED excludes any mismatched entries. Furthermore, we apply whitespace correction (Bast et al., 2023) to the NZZ ground truth, addressing issues such as merged words and unseparated punctuation marks that we identified in this dataset. In the FILTERED dataset, all Latin texts identified in the GNL dataset are removed.

4.2 Training and Test Data Description

Training data. The resulting training dataset is composed of three distinct parts. Der Diabetiker makes up 6% of the training data, the NZZ dataset contributes 56%, and the GNL dataset accounts for the remaining 38%. We vary the inclusion of the GNL portion in our experiments, as this dataset is more distant from the target data source (medical periodicals) and time period, whereas NZZ is more closely aligned with the target source and time frame.

[7] https://github.com/Shef-AIRE/llms_post-ocr_correction

	ALL-DATA	FILTERED
No. text spans	6371	4985
No. tokens	150k	118k
μ CER	0.85	0.24
σ CER	6.45	2.19

Table 1: General description of the training dataset configurations. CER statistics reflect the initial raw OCR quality

	ALL-DATA		FILTERED	
	μ CER	σ CER	μ CER	σ CER
NZZ	0.7	6.34	0.23	2.8
GNL	1.21	7.05	0.31	0.34
DD	0.03	0.09	0.03	0.09

Table 2: CER statistics for raw OCR grouped by data source and dataset configuration. DD stands for Der Diabetiker

The general description of the resulting training dataset configurations including the three datasets is provided in Table 1.

The CER statistics for raw OCR, grouped by data source and training set configuration, are presented in Table 2. It presents the average and standard deviation of the CER for each section of the training dataset, providing insight into the OCR quality across the dataset-specific training samples before and after filtering.

The initial OCR quality for Der Diabetiker is generally high, as reflected in the CER statistics for the training data shown in Table 2. To ensure a balanced representation of different error magnitudes in both the training and test sets, we examined the error categories within the Der Diabetiker data. Through this analysis, we found that approximately 7% of the data (54 out of 760 paragraphs) has a CER of 0.1 or higher, where the OCR output resulted in text spans that were significantly altered, making it nearly impossible to understand the meaning without the surrounding context. These errors occurred in pages with complex layouts and rare fonts. In contrast, 31% of the paragraphs (242 out of 760) had a CER between 0 and 0.1, with minor errors like missing umlauts, lowercase letters instead of capitals, and spacing issues. While these errors occasionally altered the meaning of some words, the overall meaning of

OCR Text	Ground Truth	CER	CER_interval
für das Arzt-Patient-Uerhältnis eher schädlich als nütz- lich sind? Hat schließlich, um auch diese Frage noch Dr. Josef Issels begrüßt vor Prozeßbeginn Staatsanwalt	für das Arzt-Patient-Verhältnis eher schädlich als nütz¬ lich sind? Hat schließlich, um auch diese Frage noch Dr. Josef Issels begrüßt vor Prozeßbeginn Staatsanwalt	0.0121951219 512195	<0.1&!=0
Abb. 33 Offnen einer Sprudelflasche Durch Änderung der Lage der Sprudelflasche strömt der Druck der senkrecht aufsteigenden Kohlensäure: a b. c. a) voll gegen die Öffnung	Abb. 33 Öffnen einer Sprudelflasche Durch Änderung der Lage der Sprudelflasche strömt der Druck der senkrecht aufsteigenden Kohlensäure: a. b. c. a) voll gegen die Öffnung	0.0117647058 823529	<0.1&!=0
'^ÄX"^ • Komplikationen und deren Vor¬ sorge beim juvenilen Diabetes • Immunsuppression	sche Aspekte. • Komplikationen und deren Vor¬ sorge beim juvenilen Diabetes • Immunsuppression	0.1494252873 563218	>=0.1

Table 3: Examples of raw OCR CER error categories - minor (<0.1&!=0) and major (>=0.1)

the text remained largely recoverable. The remaining 464 paragraphs had perfect OCR (CER = 0).

Based on these observations, we decided to use the identified error categories to balance the Der Diabetiker data in both the training and testing sets. This approach allows us to better assess model performance, particularly in terms of how well the models handle perfect OCR text (ensuring they do not degrade its quality) and how they perform with varying levels of error. The following error categories were introduced for both data balancing and further analysis:

- [NONE]: CER = 0 – perfectly recognized text
- [MINOR]: $0 < \text{CER} < 0.1$ – minor errors that do not significantly alter the text. These include issues such as missing umlauts, lowercase letters instead of capitals, and spacing errors, where the text remains recognizable and the meaning is generally preserved.
- [MAJOR]: CER >= 0.1 – substantial errors that significantly alter the text, where the meaning of the text is changed or obscured. Examples can include missing half-lines or sequences of characters that are unrecognizable due to page damage, where context is essential for comprehension. Furthermore, problems arise when the scan inadvertently includes partial text from adjacent pages.

An example of this categorization is shown in Figure 3

Test data. The test set consists of 376 paragraphs from Der Diabetiker, selected through shuffling and stratified sampling according to the CER error category. It includes 23 paragraphs with major errors, 146 with minor errors, and 207 with perfect OCR.

4.3 Finetuning Setup

As a baseline, we finetune German BART base[8] on our sequence pairs. This model is a finetuned version of *facebook/bart-base* on the German MultiLingual Summarization dataset, MLSUM (Scialom et al., 2020).

For instruction tuning of generative models, we train LoRA adapters (Hu et al., 2021) with PEFT (Mangrulkar et al., 2022) following the methodology from (Thomas et al., 2024). We use Llama 2 13b models specifically optimized to process German text[9].

Additionally, we experiment with the German Mistral 7B, which is recommended by the developers for offering a good trade-off between performance and computational efficiency. The prompt is the translation into German of the prompt from (Thomas et al., 2024). The exact prompt formulation is as follows:

```
    f"""### Anweisung:
Korrigieren Sie die OCR-Fehler
im bereitgestellten Text.

### Eingabe:
{example['OCR Text']}

### Antwort:
{example['Ground Truth']}
"""
```

We conduct finetuning using two combined scenarios:

1. A comparison between ALL-DATA, which includes mismatching spans, and manually filtered data (FILTERED).

[8] https://huggingface.co/Shahm/bart-german

[9] https://github.com/jphme/EM_German

OCR Text	Ground Truth	old_CER	CER_interval	Model Correction	new_CER	CER_reduction
von Diabetikern besonders geschätzt. Schulte-Maure! Kornbrennerei seit 1848, Castrop-Rauxel ...zpoaäfes Ce Zhre diätetischen Vabriegsmitlel 2	von Diabetikern besonders geschätzt. Schulte-Rauxel Kornbrennerei seit 1848, Castrop-Rauxel ...wo kaufen Sie Ihre diätetischen Nahrungsmittel?	0.13	>=0.1	von Diabetikern besonders geschätzt. Schulte-Rauxel Kornbrennerei seit 1848, Castrop-Rauxel ...zur Herstellung diätetischen Vollkornmittel 2	0.18	-36.8

Table 4: Post-OCR correction of an advertisement by Mistral 7b (ALL-DATA, not augmented with GNL)

2. In addition to the first setup, we compare the base training set, which includes Der Diabetiker and NZZ, with the same set augmented by GNL, denoted as [+GNL].

4.4 Evaluation Metrics

We measure average CER and WER, as well as CER and WER within error categories for the proposed training data configurations. WER is particularly critical for our data, as accurate word counts are essential for further comparisons across time periods and are also used for temporal topic modeling.

To investigate improvements in relation to the defined error categories, we assess the percentage of text spans with improved OCR quality compared to those with deteriorated or unchanged quality. This percentage is calculated as the ratio of texts with a positive CER reduction and WER reduction to the total number of texts in each error category. The CER reduction, as defined in previous work (Thomas et al., 2024), is determined using the following formula:

$$\text{CER}_{\text{reduction}} = \left(\frac{\text{CER}(gt, ocr) - \text{CER}(gt, pr)}{\text{CER}(gt, ocr)} \right) \times 100 \quad (1)$$

where *gt* denotes the ground truth, *ocr* represents the OCR output, and *pr* indicates the generative model prediction. WER reduction is calculated similarly using the corresponding WER values. To calculate WER and CER we use JiWER[10], a package for the evaluation of automatic speech recognition systems, which supports CER and WER measures. These measures are computed using the minimum edit distance between one or more reference sentences and their corresponding hypothesis sentences.

[10] https://pypi.org/project/jiwer/

	ALL-DATA		FILTERED	
	CER	WER	CER	WER
raw OCR	0.02	0.09	**0.02**	0.09
BART 140M	0.03	0.1	0.03	0.11
BART 140M [+GNL]	0.03	0.11	0.03	0.11
Mistral 7b	**0.02**	**0.07**	0.07	0.27
Mistral 7b [+GNL]	0.07	0.26	0.1	0.45
Llama-2 13b	0.25	0.28	0.03	**0.08**
Llama-2 13b [+GNL]	0.08	0.28	0.13	0.63

Table 5: Average error rate before (light-gray row) and after post-OCR correction

5 Results

5.1 Average Performance

Table 5 presents the average CER and WER across various models and dataset configurations. On average, none of the models achieves a reduction in CER. BART demonstrates stable performance across all configurations; however, it slightly increases both CER and WER, thereby deteriorating the initial OCR quality. Among the generative models, Mistral 7b stands out by maintaining CER levels and achieving a 22% reduction in WER when trained on the complete dataset without filtering (ALL-DATA).

In Table 4, we present an example of successful word correction by Mistral 7b trained on ALL-DATA and not augmented with GNL. The paragraph is categorized as a major error, as its initial raw OCR score is 0.13. The red frame highlights the OCR error that was subsequently corrected by the model, as shown in the green frame within the model correction column. The context includes the name of the location, Castrop-Rauxel, associated with the company Schulte-Rauxel. We have highlighted in blue the contextual information that could potentially assist the model in making the correction.

We identified several instances where the Mistral 7b model successfully recovered words from context. In contrast, other models, including

	ALL-DATA						FILTERED					
	[MINOR]		[MAJOR]		[ALL]		[MINOR]		[MAJOR]		[ALL]	
	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
BART 140M	26	25	26	17	26	24	29	26	35	30	29	26
BART 140M [+GNL]	25	24	30	22	25	24	25	24	39	35	27	25
Mistral 7b	**61**	**64**	39	35	**58**	**60**	51	48	30	39	48	47
Mistral 7b [+GNL]	53	56	43	43	52	55	**54**	**57**	43	**57**	**53**	**57**
Llama-2 13b	54	58	**52**	43	54	56	51	50	**48**	43	51	49
Llama-2 13b [+GNL]	53	55	43	**52**	52	55	49	50	**48**	35	48	48

Table 6: Percentage of corrected paragraphs in terms of WER and CER in each error category (%)

BART, were unable to perform similar corrections for the same paragraphs. Further investigation is required to understand the factors contributing to this difference.

We conducted a manual analysis of a subset of model outputs where a decrease in CER was observed. In several instances, the models exhibited repetition of punctuation marks and words after partially correcting the input sequence. Additionally, LLaMA 2 occasionally reproduced parts of the prompt in its output. These repetitions were not filtered out prior to metric evaluation. Further investigation is needed to better understand and mitigate these issues.

When trained on the manually filtered dataset (FILTERED), all models exhibit an increase in the average CER (Table 5). This outcome may be attributed to the reduced dataset size following the filtering process. However, Llama-2 13b demonstrates an 11% improvement in WER despite the reduction in dataset size.

On average, we observe that the inclusion of GNL data does not lead to improvements in the reduction of either CER or WER. Nevertheless, it is worth noting that GNL data might prove beneficial in addressing specific types of errors within error categories - minor or major. To explore this possibility further, we conduct a detailed analysis of models' performance within categories in the following.

5.2 Performance in Error Groups

To investigate how models perform across the error categories outlined in Section 4.2, we calculate the percentage of texts in the test set where error rates improved following post-OCR correction.

As detailed in Section 4.2, the test set comprises 23 paragraphs classified as having major errors and 146 paragraphs categorized as having minor errors. The percentage of corrected paragraphs is determined by computing the ratio of paragraphs within a given error category that exhibited a positive CER or WER reduction after post-correction (CER or WER reduction > 0) to the total number of paragraphs in that category.

We analyze this performance across three distinct categories: minor errors, major errors, and the combined category (all errors), which aggregates all instances where the initial raw OCR CER was greater than 0. Table 6 summarizes the percentage of corrected paragraphs for each model and dataset configuration, offering a comprehensive view of how effectively these models address errors across categories.

Among the models evaluated, BART demonstrated the least success in correcting paragraphs across both minor and major error categories.

Mistral 7b corrected over 60% of paragraphs with minor errors in terms of both WER and CER when trained on the ALL-DATA configuration. However, its performance dropped when dealing with more challenging errors, with the model correcting less than half of the paragraphs containing such difficult issues.

In contrast, Llama-2 demonstrated a more balanced performance across error categories. It corrected more than half of the paragraphs in terms of CER without augmentation, and over half in terms of WER when GNL augmentation was applied.

Through our manual analysis, we observed that Mistral, in particular, exhibited a certain level of creativity when handling major errors, when using the same configuration settings as the other models. This creativity was apparent in its ability to address complex error patterns, but it sometimes led to substitutions that, while contextually relevant, deviated from the exact ground truth. In these instances, Mistral was able to replace nonsensical or garbled character sequences with text that, although thematically similar, did not align perfectly with the original source.

For example, as shown in Table 4, Mistral

	ALL-DATA		FILTERED	
	ERR %	GT %	ERR %	GT %
BART 140M	**24**	72	20	60
BART 140M [+GNL]	**24**	70	**22**	62
Mistral 7b	11	90	14	84
Mistral 7b [+GNL]	12	**91**	14	**87**
Llama-2 13b	12	87	17	80
Llama-2 13b [+GNL]	13	86	17	83

Table 7: Percentage of paragraphs with unchanged error (ERR) and those with preserved perfect OCR quality (GT). The highest percentages in both columns are highlighted

corrected the misrecognized part of the paragraph, which in the ground truth should have read "... wo kaufen Sie ihre diätetischen Nahrungsmittel?" (translating to "...where do you buy your dietary foods?") by replacing it with "... zur Herstellung diätetischen Vollkornsmittel" ("...for the production of dietary whole grain products"). While both sequences are related in topic (dietary foods), the produced variations decrease the accuracy.

When we remove misaligned text spans from the dataset (in the FILTERED dataset configuration), the addition of GNL augmentation begins to show a positive impact on error correction for Mistral across both error categories. Specifically, Mistral corrects 10% more paragraphs in terms of WER and 5% more in terms of CER when GNL is included, compared to the configuration without it.

This could be attributed, in part, to the larger size of the ALL-DATA dataset, which is 27.8% larger than the FILTERED dataset. Additionally, the inclusion of misaligned passages may be enhancing Mistral 7b's ability to recover words from context. These misaligned spans could provide valuable contextual clues, aiding the model in making more accurate corrections. This potential relationship between misalignment and model performance warrants further exploration to fully understand how these factors interact and contribute to the model's effectiveness.

Interestingly, when we examine the results in the major error category for both dataset configurations, both BART and Mistral show improvements with the inclusion of GNL, demonstrating better performance in terms of both CER and WER. This suggests that the addition of GNL augmentation may help both models address more challenging errors.

We further investigate the cases with the perfect initial OCR (error-free cases) to determine which models preserve a higher proportion of accurately OCR-ed spans. In addition, we analyze spans with zero CER reduction to identify which models leave a higher percentage of errors unchanged compared to others. The results are summarized in Table 7, where GT indicates the percentage of OCR spans that perfectly match the ground truth, and ERR reflects the percentage of spans with unchanged errors. BART exhibits a higher percentage of unchanged errors compared to the generative models and preserves fewer perfectly OCR-ed spans than both Llama 2 and Mistral. In contrast, Mistral models, retain the highest proportion of accurately OCR-ed spans.

6 Conclusion

This paper compares the performance of large language models, specifically BART as an encoder-decoder, and Llama 2 13b and Mistral as generative models, for post-OCR correction of the German historical periodical Der Diabetiker, published by the German Diabetes Association. We examine the impact of different dataset configurations and the effect of dataset augmentation with data from a distant source.

The results suggest that BART detects fewer errors compared to the generative models. However, since BART does not correct these errors, it also avoids introducing larger changes — an issue that we observe in the generative models. Also, BART tends to correct more spans that were already accurate in the first place, leading to unnecessary modifications. This behavior aligns with the average CER and WER scores in Table 5, where BART shows a decline in OCR quality, but the degradation is not as severe as observed with some generative models. This could imply that BART's stability comes at the cost of detecting fewer errors overall.

Among the evaluated models, Mistral 7b stands out as the most promising in terms of performance on historical data from patient organization periodicals. It achieves a significant 22% improvement in average WER and retains the highest proportion of correctly OCR-ed paragraphs compared to other models. Despite these strengths, Mistral maintains the average CER without improvement, and further investigation is needed to understand how it handles major errors. Specifically, more research is required to manage the model's creativity in generating corrections, ensuring that it

produces more accurate and contextually relevant outputs without deviating from the ground truth.

6.1 Acknowledgements

This research is funded by the European Union (ERC, ActDisease, ERC-2021-STG 10104099). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The authors would like to thank the **Centre for Digital Humanities and Social Sciences** at Uppsala University for providing us with the computational resources for training and evaluating our models.

References

Hannah Bast, Matthias Hertel, and Sebastian Walter. 2023. Fast whitespace correction with encoder-only transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 389–399, Toronto, Canada. Association for Computational Linguistics.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017. Icdar2017 competition on post-ocr text correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1423–1428.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Atli Jasonarson, Steinór Steingrímsson, Einar Sigursson, Árni Magnússon, and Finnur Ingimundarson. 2023. Generating errors: OCR post-processing for Icelandic. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 286–291, Tórshavn, Faroe Islands. University of Tartu Library.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daniel Lopresti. 2008. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, AND '08, page 9–16, New York, NY, USA. Association for Computing Machinery.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural ocr post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2013. The impact dataset of historical document images. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, HIP '13, page 123–130, New York, NY, USA. Association for Computing Machinery.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. Icdar 2019 competition on post-ocr text correction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. BART for post-correction of OCR newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings*

of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, pages 116–121, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Martijn Veninga. 2024. LLMs for OCR Post-Correction. Master Thesis in Computer Science.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2024. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.