Evaluating Robustness of LLMs in Question Answering on Multilingual Noisy OCR Data

Bhawna Piryani
bhawna.piryani@uibk.ac.at
University of Innsbruck
Innsbruck, Austria

Jamshid Mozafari
jamshid.mozafari@uibk.ac.at
University of Innsbruck
Innsbruck, Austria

Abdelrahman Abdallah
abdelrehman.abdallah@uibk.ac.at
University of Innsbruck
Innsbruck, Austria

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle
La Rochelle, France
University of Ljubljana
Ljubljana, Slovenia

Adam Jatowt
adam.jatowt@uibk.ac.at
University of Innsbruck
Innsbruck, Austria

Abstract

Optical Character Recognition (OCR) plays a crucial role in digitizing historical and multilingual documents, yet OCR errors - imperfect extraction of text, including character insertion, deletion, and substitution can significantly impact downstream tasks like question-answering (QA). In this work, we conduct a comprehensive analysis of how OCR-induced noise affects the performance of Multilingual QA Systems. To support this analysis, we introduce a multilingual QA dataset `MultiOCR-QA`, comprising 50K question-answer pairs across three languages, English, French, and German. The dataset is curated from OCR-ed historical documents, which include different levels and types of OCR noise. We then evaluate how different state-of-the-art Large Language Models (LLMs) perform under different error conditions, focusing on three major OCR error types. Our findings show that QA systems are highly prone to OCR-induced errors and perform poorly on noisy OCR text. By comparing model performance on clean versus noisy texts, we provide insights into the limitations of current approaches and emphasize the need for more noise-resilient QA systems in historical digitization contexts.

CCS Concepts

• **Information systems → Question answering**; **Content analysis and feature selection**.

Keywords

Multilingual QA, OCR Text, Large Language Models

ACM Reference Format:
Bhawna Piryani, Jamshid Mozafari, Abdelrahman Abdallah, Antoine Doucet, and Adam Jatowt. 2025. Evaluating Robustness of LLMs in Question Answering on Multilingual Noisy OCR Data. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761295

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761295

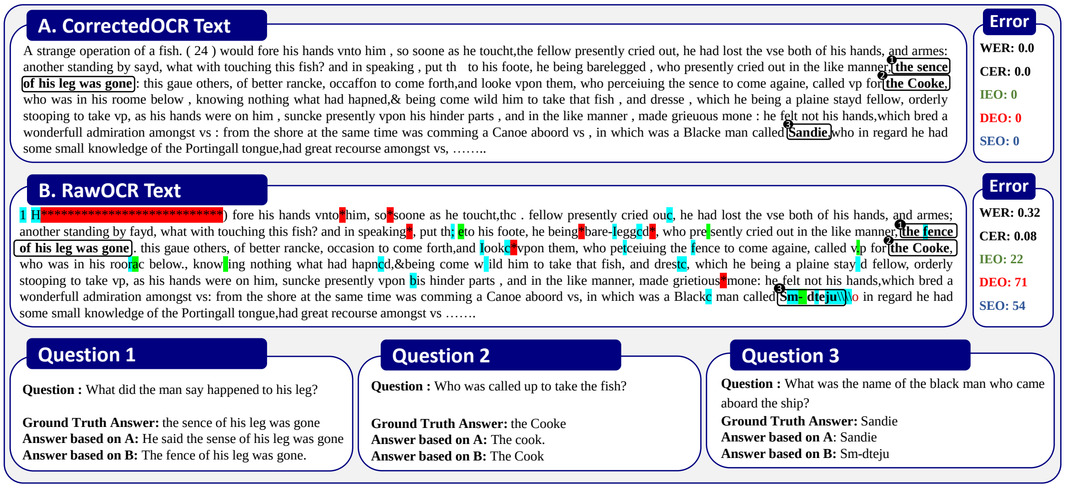
1 Introduction

Optical Character Recognition (OCR) technology has played a crucial role in digitizing and providing access to historical texts. Over the past decade, significant advancements in OCR have improved text recognition accuracy, leading to the development of large-scale digital libraries of historical texts [31]. These libraries serve as valuable resources for researchers, professionals and the general public, enabling access to old manuscripts, newspapers, and other archival materials. Historical documents hold a wealth of knowledge, offering insights into past events, cultures, and people. Many professionals such as historians, journalists, or sociologists rely on these heritage collections for various research and analysis tasks.

Historical corpora have already been used as an underlying text for numerous natural language processing (NLP) tasks, including Named Entity Recognition (NER), topic modeling, text classification, neural ranking, and understanding semantic changes in language over time [6, 7, 19, 29, 37]. While these tasks offer valuable insights, it is expected that automatic question answering (QA) systems become an important interface to the vast historical collections. QA systems enable direct and intuitive access to information, offering a powerful means of analyzing, searching and understanding historical texts. By retrieving precise answers to user queries, QA can significantly enhance the accessibility, interpretability, and utility of historical corpora, making them more easily accessible and more useful for scholars, researchers, and the general public alike.

Despite the advancements in OCR technology, significant challenges such as character misrecognition and structural inconsistencies persist. OCR-generated text, referred as **RawOCR text** in this paper, often contains errors due to the degraded conditions and non-standard character of historical documents. Factors like faded ink and paper, irregular fonts, physical damage, and printing inconsistencies cause recognition errors that negatively impact downstream NLP tasks such as information retrieval, machine translation, and QA systems. Since QA models heavily depend on the quality of the input text, errors in RawOCR text can significantly affect the accuracy and reliability of generated answers.

For instance, in German, a passage with OCR error "*Der Bericht der Lagsatzungsgesandtschaft wird verlesen undvon Hrn. Bürgermeistet Mousson als erstem Gesandten desStandes Zürich mit einigen*



A. CorrectedOCR Text

A strange operation of a fish. (24) would fore his hands vnto him , so soone as he toucht,the fellow presently cried out, he had lost the vse both of his hands, and armes: another standing by sayd, what with touching this fish? and in speaking , put th to his foote, he being barelegged , who presently cried out in the like manner, the sence of his leg was gone: this gaue others, of better rancke, occaffon to come forth,and looke vpon them, who perceiuing the sence to come againe, called vp for the Cooke, who was in his roome below , knowing nothing what had hapned,& being come wild him to take that fish , and dresse , which he being a plaine stayd fellow, orderly stooping to take vp, as his hands were on him , suncke presently vpon his hinder parts , and in the like manner , made grieuous mone : he felt not his hands,which bred a wonderfull admiration amongst vs : from the shore at the same time was comming a Canoe aboord vs , in which was a Blacke man called Sandie,who in regard he had some small knowledge of the Portingall tongue,had great recourse amongst vs, ……..

Error

WER: 0.0

CER: 0.0

IEO: 0

DEO: 0

SEO: 0

B. RawOCR Text

1 H***************************) fore his hands vnto*him, so*soone as he toucht,thc . fellow presently cried ouc, he had lost the vse both of his hands, and armes; another standing by fayd, what with touching this fish? and in speaking*, put th; eto his foote, he being*bare-Ieggcd*, who pre sently cried out in the like manner, the fence of his leg was gone. this gaue others, of better rancke, occasion to come forth,and Iookc*vpon them, who petceiuing the fence to come againe, called v.p for the Cooke, who was in his roorac below., know ing nothing what had hapncd,&being come w ild him to take that fish, and drestc, which he being a plaine stay d fellow, orderly stooping to take vp, as his hands were on him, suncke presently vpon bis hinder parts , and in the like manner, made grietious*mone: he felt not his hands,which bred a wonderfull admiration amongst vs: from the shore at the same time was comming a Canoe aboord vs, in which was a Blackc man called Sm- dteju\\\o in regard he had some small knowledge of the Portingall tongue,had great recourse amongst vs …….

Error

WER: 0.32

CER: 0.08

IEO: 22

DEO: 71

SEO: 54

Question 1

Question : What did the man say happened to his leg?

Ground Truth Answer: the sence of his leg was gone
Answer based on A: He said the sense of his leg was gone
Answer based on B: The fence of his leg was gone.

Question 2

Question : Who was called up to take the fish?

Ground Truth Answer: the Cooke
Answer based on A: The cook.
Answer based on B: The Cook

Question 3

Question : What was the name of the black man who came aboard the ship?
Ground Truth Answer: Sandie
Answer based on A: Sandie
Answer based on B: Sm-dteju

Figure 1: An example of CorrectedOCR and RawOCR text from the MultiOCR-QA dataset for the English language, highlighting different types of errors along with questions corresponding to this text. WER and CER denote Word Error Rate and Character Error Rate, respectively, indicating the level of errors in the text. The green highlights represent insertion errors, where IEO denotes Insertion Edit Operations - the number of insertions needed to transform RawOCR into CorrectedOCR. Similarly, red and blue highlights indicate deletion and substitution errors, with DEO and SEO representing Deletion Edit Operations and Substitution Edit Operations, respectively. The black boxes with numbers in the CorrectedOCR and RawOCR text correspond to the answers for each question in the paragraph.

Bemerkungen begleitet.[1]" contains multiple insertion, deletion, and substitution errors, such as "*Lagsatzungsgesandtschaft*" (should be "Tagsatzungsgesandtschaft") and *"Bürgermeistet"* (should be "Bürgermeister"). When a QA model encounters such a noisy OCR text, it may generate an incorrect answer. For example, given the question: "*Wer hat den Bericht der Tagsatzungsgesandtschaft verlesen?*" (Who read the report of the parliamentary delegation?) the model incorrectly responds "*Der Bericht der Tagsatzungsgesandtschaft wurde von Hrn. Bürgermeistet Mousson verlesen.*" (The report of the parliamentary delegation was read out by Mr. Mayor Mousson.) which retains the OCR error and potentially misleads the QA system. This example highlights how even minor OCR errors can significantly affect QA accuracy, resulting in misleading or incorrect answers.

Although extensive research has focused on improving OCR accuracy and post-processing correction techniques, the specific impact of OCR noise on QA remains largely unexplored. Previous studies have examined challenges related to OCR in information retrieval (IR) [3], historical text processing [14, 24], and entity recognition [10]. However, a systematic investigation of how OCR errors affect QA model performance is still missing. Furthermore, our study contributes also to the body of work on robustness of LLMs against noisy inputs such as noisy prompts [35], however we do it in a quite novel yet realistic setting.

In this paper, we address this gap in understanding how LLMs perform in QA tasks when dealing with noisy OCR-generated text. In particular, we first introduce `MultiOCR-QA`[2], a new multilingual QA dataset covering English, French, and German. This dataset includes both **RawOCR text** (OCR-generated text with errors) and **CorrectedOCR text** (ground truth), allowing for a direct comparison of QA performance under different text quality conditions. To generate contextually relevant question-answer pairs from historical text excerpts, we leverage instruction-fine-tuned LLMs. We then systematically evaluate the impact of different types of OCR errors—insertions, deletions, and substitutions—on QA model performance, offering new insights into the strengths and limitations of LLMs when processing noisy historical data. Figure 1 illustrates an example from the `MultiOCR-QA` dataset with various OCR errors impacting text accuracy. It presents QA pairs and the corresponding responses for questions when using CorrectedOCR and RawOCR text as input context to Gemma 2–27b, demonstrating the effects of OCR noise on QA performance. For instance, *Question 1: "What did the man say happened to his leg?"* produces different answers depending on whether clean or noisy text was input. When using the CorrectedOCR context, the model produces the response: *"He said his sense of the leg was gone."* In contrast, the response based on the RawOCR context is: *"The fence of his leg is gone,"* demonstrating how OCR-induced noise can significantly alter the meaning of the answer and degrade the QA accuracy.

In summary, we make the following contributions in this work:

- We introduce `MultiOCR-QA`, a new multilingual QA dataset from historical texts in English, French, and German, featuring both raw and corrected OCR text, that allows direct comparison under varying text quality conditions.
- Using `MultiOCR-QA`, we conduct a comprehensive evaluation of LLM robustness against noisy OCR text analyzing and quantifying their impact on QA performance.

[1] English Translation: The report of the legislative mission is read out and accompanied by a few comments from Mr. Mayor Mousson, the first envoy of the Zurich state.
[2] The dataset is available at https://github.com/DataScienceUIBK/MultiOCR-QA

- We categorize different types and degrees of OCR errors for the purpose of evaluating their individual impact on QA performance. This allows providing comprehensive insights into how LLMs handle OCR-related challenges for different types and severity of errors in each of the studied languages.

2 Related Work

Several research works have been carried to study the limitations of OCR for historical documents and its impact on information retrieval (IR). Croft et al. [2] and Traub et al. [32] examined how OCR errors reduce retrieval effectiveness. Chiron et al. [1] found that 7% of the relevant documents were missed due to OCR misrecognition, demonstrating the risk of failure in matching noisy texts to user queries. While these studies highlight OCR challenges, they focus primarily on document retrieval and not on question answering (QA), which requires a more fine-grained understanding of text.

Beyond IR, OCR errors have been studied in multiple tasks, including named entity recognition (NER) [11, 13], entity linking [20], text classification [34, 41], topic modeling [23, 40], document summarization [16], machine translation [8, 17], and document ranking [9]. OCR noise has been shown to significantly degrade performance across these tasks. For instance, van Strien et al. [33] demonstrated that low-quality documents negatively impact multiple tasks, including dependency parsing and sentence segmentation. Hamdi et al. [12] found that 80.75% of the named entities were misrecognized due to OCR errors, causing substantial drops in accuracy. Similarly, Hamdi et al. [13] reported that the F1-score for NER drops from 90% to 50% when the character error rate increases from 2% to 30%. In topic modeling, Mutuvi et al. [23] showed that OCR noise distorts the identification of key topics. For document retrieval, de Oliveira et al. [3] analyzed performance degradation at different OCR error rates, noting that retrieval effectiveness begins to decline at a word error rate of 5% and worsens as the error rate increases. Giamphy et al. [9] further examined the impact of different types of OCR noise on document ranking and advocated for developing more robust ranking methodologies.

Despite these insights into OCR's effects on IR and NLP tasks, research on its impact on question answering remains limited. In the context of historical document collections, the only existing QA dataset, ChroniclingAmericaQA [25], focuses primarily on creating a QA dataset from historical newspapers rather than systematically analyzing how different types of OCR errors affect QA performance. While studies on document retrieval and IR highlight OCR-related challenges, a comprehensive investigation into QA performance under different types and severity levels of OCR errors is still missing. Our work fills this gap by introducing a multilingual QA dataset (`MultiOCR-QA`) and providing a detailed evaluation of large language models (LLMs) on the RawOCR text of `MultiOCR-QA`.

3 Methodology

To systematically investigate the impact of OCR errors on QA systems, we constructed `MultiOCR-QA`, a new multilingual QA dataset derived from historical texts processed with OCR. This section details the two main stages of the dataset creation pipeline: Data Collection and Question-Answer Generation. Figure 2 provides an overview of this process.

3.1 Data Collection

We first describe the process of collecting documents to generate question-answer pairs for our study. Although several historical text datasets exist, such as the IMPACT Project[3] and the Europeana Newspaper Project[4], which include digitized images of historical documents alongside their OCR-processed text, these lack systematically aligned ground truth corrections for a robust analysis of OCR noise effects. The ICDAR 2019 POST-OCR Text Correction dataset[5] [27] is an exception here as it provides both RawOCR text along with its aligned ground truth (called CorrectedOCR), making it especially suitable for our research objectives. Furthermore, several English, French, and German document sources included in the IMPACT and Europeana projects are already incorporated into the ICDAR 2019 dataset, offering a diverse collection of cleaned resources for historical OCR-based studies.

ICDAR 2019 dataset includes over 22 million OCR-processed characters and their corresponding, aligned ground truth across several European languages. We focus on English, French, and German languages due to several reasons. First, French and German are among the most well-represented languages in the ICDAR 2019 dataset regarding their document numbers. Second, all three languages are high-resource and well-supported by the current large language models (LLMs), which increases the reliability of QA generation and evaluation [18]. Since noisy OCR text is expected to pose already a significant challenge for LLMs, incorporating low-resource languages, where LLM performance tends to be generally weaker, might introduce additional complexity. Third, each of these languages has an existing, publicly available QA dataset like SQuAD [26] for English, FQuAD [4] for French, and GermanQuAD [22] for German, which enable effective instruction fine-tuning of LLMs for high-quality QA pair generation. Therefore, we limited our study to these three languages, intending to expand to low-resource and typologically diverse languages in the future work.

***Language Specific Data Collection*:** The texts in ICDAR 2019 dataset originally came from various historical document repositories.

- `English`: The documents for the English language in the ICDAR 2019 dataset are sourced from IMPACT - British Library, comprising a total of 150 files.
- `French`: For the French language, the ICDAR 2019 dataset provides a collection of 2,800 files obtained from three sources: the HIMANIS[6] Project, IMPACT - National Library of France, and the RECEIPT[7] dataset.
- `German`: The German-language dataset includes the OCR-processed text from multiple sources, such as, front pages of the Swiss newspaper NZZ[8], IMPACT - German National Library, GT4Hist-dta19 dataset, GT4Hist - EarlyModern-Latin, GT4Hist - Kallimachos, GT4Hist - RefCorpus-ENHG-Incunabula, and GT4Hist - RIDGES-Fraktur[9] [28]. The German dataset in ICDAR 2019 originally contained 10,032 files.

[3] https://www.digitisation.eu/impact-dataset/
[4] http://www.europeana-newspapers.eu/
[5] https://sites.google.com/view/icdar2019-postcorrectionocr
[6] https://www.himanis.org
[7] http://findit.univ-lr.fr/
[8] https://zenodo.org/records/3333627
[9] https://zenodo.org/records/1344132

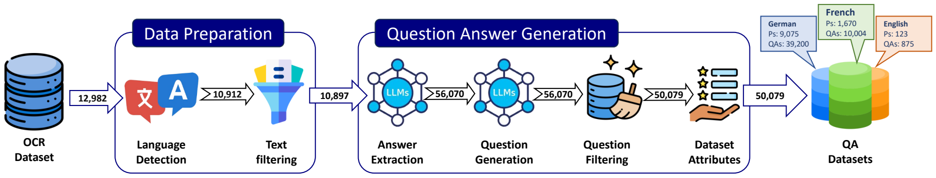


Figure 2: The Pipeline for `MultiOCR-QA` generation: Arrows represent the output quantity based on the number of documents and callouts illustrate the statistics for English, French and German Language. Ps and QAs denote the number of paragraphs and question-answer pairs, respectively.

***Language Verification and Filtering*:** Prior to QA pair generation, we preprocessed the ICDAR dataset to ensure that each document contains text in the correct target language. We applied langdetect[10] library to detect the language of documents. The analysis revealed that some documents labeled as English, French and German were actually in other languages, particularly Latin. To maintain dataset integrity, we removed non-target language documents, resulting in the following reductions: We removed non-English documents, reducing the number of documents in the English dataset to 141. Similarly, we discarded non-French documents, reducing the dataset to 1,713 French-language files, and eliminating 1,086 Latin-language files. Finally, for German, we removed Latin or other non-German files and retained a total of 9,075 German-language files.

Furthermore, the ICDAR dataset, originally intended for post-OCR correction, contained special alignment symbols (e.g., `@`, #) to map the RawOCR text to its ground-truth counterpart. We removed these symbols from the ground-truth text before generating questions. We also excluded files where the ground-truth text was missing, resulting in the removal of 16 files for English, 3 files for French, and none for German. This preprocessing step ensured that all QA pairs were generated from text that had both CorrectedOCR text and RawOCR text.

3.2 Question-Answer Generation

To construct the multilingual QA dataset, we opted for automatic QA pair generation, as manual dataset creation would require substantial human resources. To achieve this, we instruction fine-tuned a pretrained LLM for each target language to generate QA pairs from CorrectedOCR text.

While LLMs are pretrained on diverse NLP tasks, they typically generate a variety of question types, including non-factoid and open-ended questions. Since our goal is to develop a factoid QA dataset, we fine-tuned the models using language-specific QA datasets to ensure the generation of structured, precise, and factual question-answer pairs.

Instruction fine-tuning enhances both the capabilities and controllability of LLMs [39]. Fine-tuning instruction-based datasets across multiple languages allows the model to generalize across different question-answering styles, ensuring that the generated questions remain relevant even when dealing with language-specific variations. We opted to finetune LLaMa-3.1-70B instruct model [5] separately for each language using widely adopted QA datasets.

For fine-tuning the model in English, we used the SQuAD v1 dataset [26]. We randomly selected 2,067 paragraphs and 10,570 questions from the development set and 3,000 paragraphs and 13,894 questions from the test set.

Similarly, for the French language, we fine-tuned the model on the FQuAD dataset [4], utilizing both the validation and training sets, comprising 5,689 paragraphs and 23,919 questions. For the German language, we fine-tuned the model on the GermanQuAD dataset [22] using the training and validation sets, which consisted of 3,014 paragraphs and 13,722 questions. This fine-tuning step ensured that the model accurately generated factoid-style QA pairs, reducing instances of open-ended or conversational questions.

After instruction fine-tuning, we used the fine-tuned model for each language to generate questions for the preprocessed dataset prepared in the initial step. To generate high-quality QA pairs, we employed a **two-step prompt-based approach:** Answer Extraction and Question Generation for the Extracted Answers.

Answer Extraction: The model was first prompted to extract multiple candidate answer spans from a given passage. These spans included entities, numbers, dates, locations, and key phrases that could serve as factual answers. The following prompt was used for the extraction of the candidate answer.

English Answer Extraction Prompt

System Prompt: *You are an expert at extracting key information from text. Your goal is to identify spans of text that are likely to serve as answers to potential questions based on the input passage. Focus on meaningful, distinct, and diverse snippets such as entities, nouns, verbs, adjectives, numbers, dates, and phrases. Avoid redundancy and ensure the answers are diverse, representing key information in the passage.*
User Prompt*: Given the passage below, extract several candidate spans that are likely to be answers to potential questions. Write only the extracted answers, separated by a semicolon (;). Passage : {context}*

After generating the candidate answers, we checked for duplicated answer spans. If the answers were duplicate, we removed them and retained only unique answers. This prompt was applied uniformly across all three languages, with translations adapted to each language.

Question Generation: Following answer extraction, the extracted answer spans were re-fed into the model, and it was prompted to generate questions that align with each answer while maintaining contextual relevance. The generated questions were structured

[10] https://github.com/Mimino666/langdetect

to be standalone, well-formed, and factually grounded in the passage. The following Prompt was used for Question Generation from the Extracted Answers.

> **English Question Generation Prompt**
>
> ***System Prompt:*** *You are an expert at generating standalone questions based on the provided passage. Your goal is to create a clear, relevant, and self-contained question that aligns with the information in the passage. The question should not explicitly reference the passage or require additional information to be understandable. Ensure the question is concise, well-structured, and meaningful.*
> ***User Prompt:*** *Based on the passage below, please generate a question that is relevant to the information provided. The question should be standalone, clear, and understandable without referencing the passage directly. The answer to the question should be [answer]. Passage : {context}*

Using this approach, we generated 941 questions for English, 10,522 questions for French, and 44,607 questions for German.

Dataset Filtering: After generating the dataset for each language, we applied additional filtering steps to ensure quality and consistency. Specifically, we removed questions that did not end with a question mark, duplicate questions, and questions with excessively long answers. Since LLM-generated datasets may contain hallucinated long answers, we applied this additional filtering by removing excessively long answers. As a result, we removed 66 questions for English Language, 530 questions for French Language, and 7,210 questions for German Language. This final filtering step ensured that the MultiOCR-QA dataset consisted of concise, well-structured, and factually accurate question-answer pairs.

4 Dataset Analysis

After applying all the filtering steps, we obtained the final dataset, comprising 50,079 question-answer pairs. The dataset statistics, including average paragraph length, question length, and answer length, are presented in Table 1.

4.1 Quantifying and Filtering Noise

To assess the impact of OCR noise on QA quality, we quantified the noise level in the RawOCR text using two standard metrics: Character Error Rate (CER) and Word Error Rate (WER). CER measures the proportion of character-level errors in the RawOCR text compared to the ground truth text. It is computed as the number of insertions, deletions, and substitutions (including spaces) required to transform the RawOCR text into its correct form. WER quantifies word-level discrepancies, representing the proportion of words that require modifications (insertions, deletions, or substitutions) to match ground truth. Both CER and WER were computed using the Levenshtein distance [21], which determines the minimum number of edits needed to correct the OCR-generated text. A high CER but low WER suggests that errors are concentrated within a few words (e.g., spelling variations), whereas a high WER indicates distortions across multiple words, significantly affecting readability.

Outlier Detection: To ensure a reliable analysis, we applied the Interquartile Range (IQR) method to detect and remove outliers in CER values. Outliers were defined as CER values below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR, where Q1 and Q3 represent the 33rd and 66th percentiles of the CER distribution, respectively, and IQR is the difference between Q3 and Q1. This filtering resulted in the removal of 25 English, 351 French, and 2,423 German paragraphs.

Table 1: Basic statistics of the MultiOCR-QA dataset, including question-answer (QA) pair count, paragraph count, and average text lengths across languages.

	English	French	German
#QA pairs	875	10,004	39,200
#Paragraphs	123	1,670	9,075
Avg. CorrectedOCR paragraph length (words)	271.73	297.53	217.33
Avg. RawOCR paragraph length (words)	263.46	335.73	193.23
Avg. question length (words)	8.60	8.73	8.08
Avg. answer length (words)	2.93	3.12	5.63
Avg. questions per paragraph	7.11	5.99	4.32

Table 2: OCR Error Statistics across Languages

Metric	**English**	**French**	**German**
Character Error Rate (CER)			
Mean	0.1245	0.0519	0.2816
Median	0.0729	0.0440	0.2592
Word Error Rate (WER)			
Mean	0.3047	0.1904	0.8713
Median	0.2711	0.1760	0.8730
Edit Operations			
Mean Substitutions	54.97	31.51	240.23
Median Substitutions	38.00	21.00	230.50
Mean Deletions	64.42	41.09	85.09
Median Deletions	18.00	25.00	39.00
Mean Insertions	66.91	20.90	82.60
Median Insertions	27.50	11.00	80.00

Following outlier removal, we categorized the remaining paragraphs into three noise levels based on CER percentiles for each language. Documents with CER below the 33rd percentile were classified as "low noise," those between the 33rd and 66th percentiles as "medium noise", and those above the 66th percentile as "high noise." The specific CER thresholds for each category and language were as follows:

- English: Low: CER < 0.0576, Medium: $0.0576 \leq$ CER < 0.1238, High: CER ≥ 0.1238
- French: Low: CER < 0.0357, Medium: $0.0357 \leq$ CER < 0.0558, High: CER ≥ 0.0558
- German: Low: CER < 0.2453, Medium: $0.2453 \leq$ CER < 0.2757, High: CER ≥ 0.2757

In addition to CER-based classification, we analyzed the distribution of three specific OCR error types: insertions, deletions, and substitutions. Each error type was categorized separately using a percentile-based approach, allowing for a more detailed examination of the nature and severity of OCR distortions. To further investigate OCR noise patterns, we classified insertion, deletion, and substitution errors into low, medium, and high noise levels. As illustrated in Figure 3, the distribution of these error types varies significantly across languages, reflecting differences in OCR quality and text processing challenges in English, French, and German. Additionally, Table 2 presents the statistical characteristics of the distribution of OCR error metrics across languages, including CER, WER, and edit operations, providing further insights into the OCR noise characteristics.

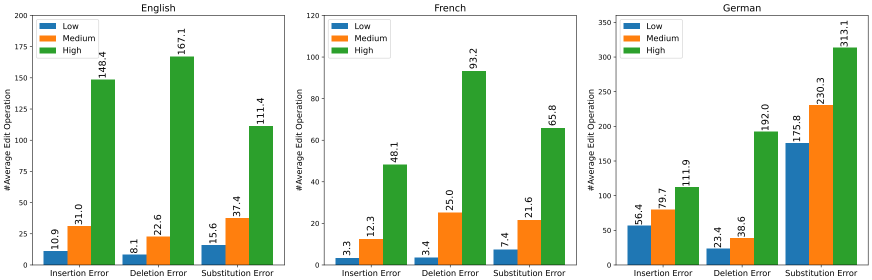


Figure 3: Statistics of insertion, deletion, and substitution errors for each language, categorized into low, medium, and high noise levels.

5 Experiments and Results

In this section, we conduct a comprehensive analysis of MultiOCR-QA from several perspectives. First, we evaluate the performance of MultiOCR-QA across various LLMs, comparing different model families and sizes to assess their effectiveness in handling OCR-generated text. Second, we investigate the impact of OCR errors on QA performance, focusing on three main error types: insertion errors, deletion errors, and substitution errors.

5.1 Experimental Settings

We conducted experiments using multiple large language models (LLMs), including Qwen2.5 7B [36], LLaMa 3.1-8B [5], Gemma-2-27B [30], Mixtral 8x22B [15], LLaMA 3.3-70B [5], and Qwen2.5 72B [36]. These models span different architectures and parameter sizes, allowing for a comprehensive comparison on OCR text.

Traditionally, QA systems are evaluated using Exact Match (EM). However, these metrics can be insufficient for LLMs, as models often generate verbose responses, leading to low EM scores even when the correct answer is included in the response. To address this limitation, we evaluate MultiOCR-QA using **BERTScore** [38] alongside **EM**. Additionally, we introduce another evaluation metric, **Contains**, to better assess MultiOCR-QA performance. Contains measures the extent to which the ground truth is present in the response generated by the model, regardless of the verbosity.

We will apply these metrics to evaluate the QA results on RawOCR texts and then on CorrectedOCR texts used as context.

5.2 Experimental Results

Tables 3, 4, and 5 present the impact of OCR errors on the performance of various LLMs in question-answering tasks for English, French, and German texts.

English text: Table 3 presents the performance of LLMs on English text using both CorrectedOCR (CP) and RawOCR (RP) paragraphs. Across all models, the transition from CP to RP negatively impacts performance. The best-performing model on CP is LLaMA-3.3-70B with a BERTScore of 66.94, followed by Gemma-2 27B at 63.99. When switching to RP, LLaMA-3.3-70B still achieves the highest BERTScore 62.87, but with a 6.08% drop, highlighting its robustness. The lowest-performing model is Mixtral 8x22B, showing the most significant impact of OCR errors. LLaMA-3.3-70B model achieves the best performance in CP for the Contains metric of 63.90, indicating its strong retrieval ability in clean text. However, it experiences a 21.3% drop in RP, suggesting moderate sensitivity to OCR errors. The lowest performing model in RP is again Mixtral 8x22B for the Contains metric, which sees a 15.70% decrease, indicating its greater vulnerability to noisy text. The EM metric shows the steepest decline across models, emphasizing that OCR errors severely impact the models' ability to generate precise answers.

French text: In Table 4 we summarize how LLMs perform on QA over French text. Models consistently perform better on CorrectedOCR, confirming and quantifying the negative influence of OCR errors on the QA accuracy. Gemma-2 27B model achieves the highest BERTScore of 76.51 on CP, demonstrating its strong ability to capture semantic similarity. Despite the 1.46% drop, it still maintains the highest performance with a BERTScore of 75.39 on RP, indicating robustness to OCR noise. Mixtral 8x22B model shows the smallest drop of 1.05%, but its overall score remains lower than the ones for the other models. For Contains, Qwen-2.5 72B achieves the highest score on CP 57.55, highlighting superior retrieval performance on clean text. However, it experiences a 19.90% drop in RP, reinforcing its vulnerability to OCR errors. LLaMA-3.1 8B also struggles, with a 19.63% decline, showing difficulty in retrieving spans from noisy text. In terms of EM, Gemma-2 27B outperforms all models with an EM of 17.46 on CP. Although it drops by 15.23%, it still maintains the best EM value 14.80 in RP. Mixtral 8x22B model is the most affected in EM, dropping by 27.99%, suggesting that OCR noise drastically reduces its accuracy in generating exact answers.

German text: Table 5 focusing on German language shows the most significant performance decline among the three languages. Unlike English and French, German exhibits the largest performance drop due to its lower OCR quality, and the models struggle more with its linguistic structure. Among the models evaluated, Gemma-2 27B achieves the highest BERTScore of 67.07 on CP, confirming its strong ability to capture semantic similarity in clean text. It also maintains the best performance on RP (63.78); however, it still experiences a 4.91% decrease, highlighting the adverse effects of OCR errors. In contrast, LLaMA-3.1 8B and LLaMA-3.3 70B show the biggest BERTScore drop (8.74% and 8.43% respectively), suggesting that these models struggle more with OCR noise. For the Contains metric, LLaMA-3.3 70B achieves the highest Contains score on CP, making it the most effective for retrieving relevant information in clean text. However, all models suffer a severe drop in Contains when moving to RP, with reductions exceeding 65%. Qwen-2.5 7B has the worst drop 66.40%, indicating that it faces challenges to

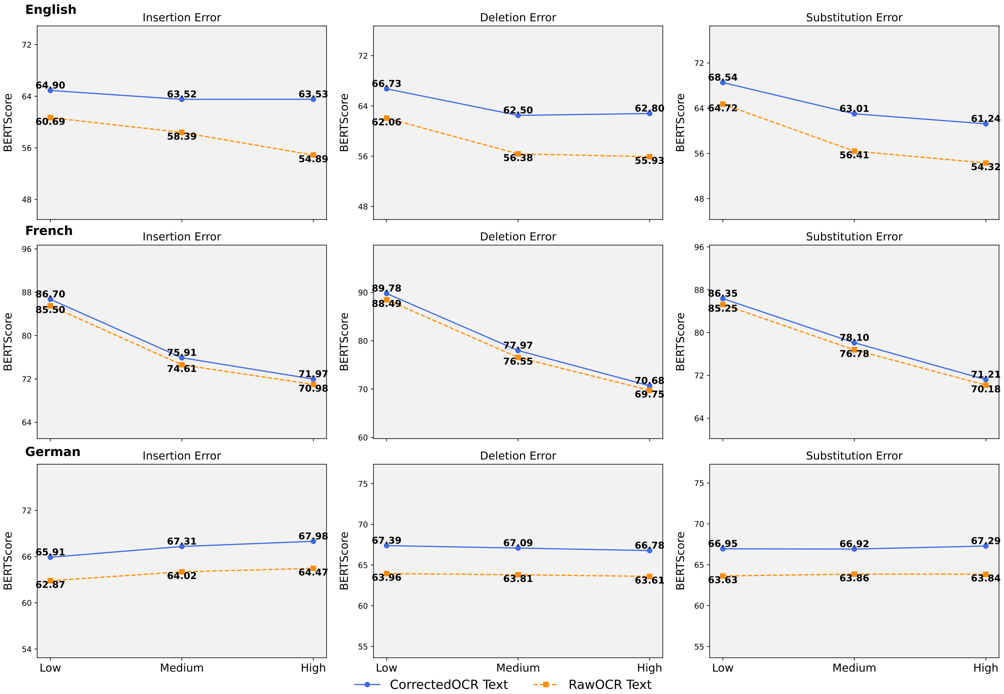


Figure 4: BERTScore of different error types on Low, medium and High categories for each Language in `MultiOCR-QA` dataset.

Table 3: Performance of LLMs on English Language: Comparison of CorrectedOCR (CP) and RawOCR Paragraphs (RP) as Context. Red numbers indicate the percentage decrease in performance with RP. Bold values highlight the highest performance for each metric with CP, while underlined values denote the best performance for each metric with RP.

Model	Parameter	BERTScore	Contains	EM
Qwen-2.5 (CP)	7B	53.28	60.04	5.21
Qwen-2.5 (RP)	7B	50.12 (5.93%↓)	45.95 (23.46%↓)	3.67 (29.56%↓)
LLaMA-3.1 (CP)	8B	55.39	63.71	0.97
LLaMA-3.1 (RP)	8B	52.18 (5.80%↓)	51.74 (18.78%↓)	0.39 (59.79%↓)
Gemma-2 (CP)	27B	63.99	58.09	**16.76**
Gemma-2 (RP)	27B	58.08 (9.23%↓)	46.20 (20.48%↓)	9.36 (44.15%↓)
Mixtral (CP)	8x22B	45.83	57.72	0.77
Mixtral (RP)	8x22B	44.72 (2.42%↓)	48.65(15.70%↓)	0.58 (24.68%↓)
LLaMA-3.3 (CP)	70B	**66.94**	**63.90**	11.39
LLaMA-3.3 (RP)	70B	62.87(6.08%↓)	50.39 (21.3%↓)	8.11 (28.80%↓)
Qwen-2.5 (CP)	72B	53.69	63.71	7.92
Qwen-2.5 (RP)	72B	50.53 (5.88%↓)	50.19 (21.24%↓)	4.44 (43.94%↓)

retrieve information from noisy text. In terms of EM, Gemma-2 27B achieves the highest EM score on CP 2.69, while Mixtral 8x22B has the lowest EM 0.078. EM scores drop drastically across all models, the largest decrease is 79.04% for Gemma-2, reinforcing that word-level distortions from OCR errors make exact answer matching nearly impossible. The Mixtral model drops 72.74% in EM, making it highly unreliable for exact answers in noisy OCR text.

Table 4: Performance of LLMs on French Language: Comparison Using CorrectedOCR (CP) and RawOCR Paragraphs (RP) as Context. Red numbers indicate the percentage decrease in performance with RP. Bold values highlight the highest performance for each metric with CP, while underlined values denote the best performance for each metric with RP.

Model	Parameter	BERTScore	Contains	EM
Qwen-2.5 (CP)	7B	73.12	54.82	11.38
Qwen-2.5 (RP)	7B	72.03 (1.49%↓)	42.99 (21.58%↓)	9.59 (15.75%↓)
LLaMA-3.1 (CP)	8B	70.16	51.35	0.64
LLaMA-3.1 (RP)	8B	69.30 (1.23%↓)	41.27 (19.63%↓)	0.61 (5.56%↓)
Gemma-2 (CP)	27B	**76.51**	52.58	**17.46**
Gemma-2 (RP)	27B	75.39 (1.46%↓)	42.05 (20.02%↓)	14.80 (15.23%↓)
Mixtral (CP)	8x22B	68.65	48.32	0.30
Mixtral (RP)	8x22B	67.93 (1.05%↓)	38.96 (19.37%↓)	0.21 (27.99%↓)
LLaMA-3.3 (CP)	70B	72.50	54.00	4.41
LLaMA-3.3 (RP)	70B	71.42 (1.49%↓)	43.22 (19.96%↓)	3.89 (11.72%↓)
Qwen-2.5 (CP)	72B	73.26	**57.55**	10.62
Qwen-2.5 (RP)	72B	71.98 (1.75%↓)	46.10 (19.90%↓)	7.84 (26.16%↓)

Summary of findings: OCR errors consistently degrade the performance of the models in English, French and German texts, resulting in maximum drop of 9.23%, 1.75%, and 8.74% in BERTScore, respectively. The most severe impact was observed in German due to the lower quality of the OCR and the complex linguistic structure. While larger models like Gemma-2 27B and LLaMA-3.3 70B demonstrate greater resilience, all models suffer substantial declines in Contains and Exact Match (EM) metrics, highlighting their weakness in retrieving and generating precise answers from noisy

Table 5: Performance of LLMs on German Language: Comparison Using CorrectedOCR (CP) and RawOCR Paragraphs (RP) as Context. Red numbers indicate the percentage decrease in performance with RP. Bold values highlight the highest performance for each metric with CP, while underlined values denote the best performance for each metric with RP.

Model	Parameter	BERTScore	Contains	EM
Qwen-2.5 (CP)	7B	62.76	15.88	0.389
Qwen-2.5 (RP)	7B	59.52 (5.16%↓)	5.33 (66.40%↓)	0.192 (50.45%↓)
LLaMA-3.1 (CP)	8B	63.87	15.36	0.457
LLaMA-3.1 (RP)	8B	58.29 (8.74%↓)	5.31 (65.37%↓)	0.135 (70.31%↓)
Gemma-2 (CP)	27B	**67.07**	11.56	**2.691**
Gemma-2 (RP)	27B	63.78 (4.91%↓)	4.47 (61.27%↓)	0.564 (79.04%↓)
Mixtral (CP)	8x22B	60.87	11.23	0.078
Mixtral (RP)	8x22B	58.14 (4.48%↓)	4.56 (59.37%↓)	0.021 (72.74%↓)
LLaMA-3.3 (CP)	70B	63.82	**17.68**	0.553
LLaMA-3.3 (RP)	70B	58.44 (8.43%↓)	6.24 (64.72%↓)	0.203 (63.23%↓)
Qwen-2.5 (CP)	72B	63.25	16.43	0.699
Qwen-2.5 (RP)	72B	60.04 (5.08%↓)	6.50 (60.44%↓)	0.167 (76.02%↓)

text. Gemma-2 27B consistently outperforms others, maintaining the highest BERTScore and EM across all languages, but still experiences notable degradation in noisy conditions. Mixtral 8x22B emerges as the most vulnerable, exhibiting the lowest performance and struggling particularly with exact answer generation.

5.3 Performance based on Different Error Types

In this section, we conduct an in-depth analysis of the impact of different types of OCR errors: insertion, deletion, and substitution on QA systems. We use Gemma-2 (27B) for this analysis, as it consistently outperform the other models across English, French, and German. As detailed in Section 4.1, each error type has been categorized into three levels: Low, Medium, and High, where Low represents minimal presence and High indicates the most frequent occurrence of a particular type of error. We evaluated MultiOCR-QA's performance across these categories, as illustrated in Figure 4.

Insertion errors introduce extraneous characters or words, leading to moderate performance degradation. At low and medium insertion levels, the effect on BERTScore remains relatively minor, suggesting that small insertions do not always disrupt semantic meaning. However, at high insertion levels, performance drops sharply, indicating that excessive insertions impair both readability and semantic coherence.

Deletion errors impact sentence coherence and factual consistency, especially when they corrupt or remove key words or essential contextual phrases. Although the impact is less pronounced at lower levels, it escalates sharply when the frequency of deletions increases. At higher levels of deletion error, the degradation in BERTScore is similar to that seen with substitution errors, highlighting how missing characters or words disrupt structured text. Substitution errors exhibit the most severe impact on QA performance in English and French, causing the steepest decline in BERTScore as their frequency increases. Since these errors modify characters within words, they often alter word meaning and disrupt sentence structure, making them highly detrimental to text comprehension.

Table 6: Performance metrics of QA systems on English text using pre-processed and post-processed OCR input. Red values indicate the percentage drop in performance compared to using Corrected Paragraph (Ground Truth) as context.

Approach	BERTScore	Contains	EM
Corrected Paragraph (Ground Truth)	63.65	57.60	16.20
RawOCR Paragraph	57.85 (9.11%↓)	45.60 (20.83%↓)	9.00 (44.44%↓)
LLM Corrected Paragraph	59.14 (7.09%↓)	42.80 (25.69%↓)	12.20 (24.69%↓)
RawOCR Corrected Answer	56.10 (11.86%↓)	41.33 (28.25%↓)	9.16 (43.46%↓)

Table 7: Performance metrics of QA systems on French text using pre-processed and post-processed OCR input. Red values indicate the percentage drop in performance compared to using Corrected Paragraph (Ground Truth) as context.

Approach	BERTScore	Contains	EM
Corrected Paragraph (Ground Truth)	65.17	52.40	14.40
RawOCR Paragraph	63.01 (3.31%↓)	41.40 (20.61%↓)	12.40 (13.89%↓)
LLM Corrected Paragraph	57.78 (11.33%↓)	29.60 (43.51%↓)	5.40 (62.50%↓)
RawOCR Corrected Answer	51.72 (20.64%↓)	19.00 (63.74%↓)	1.00 (93.06%↓)

Table 8: Performance metrics of QA systems on German text using pre-processed and post-processed OCR input. Red values indicate the percentage drop in performance compared to using Corrected Paragraph (Ground Truth) as context.

Approach	BERTScore	Contains	EM
Corrected Paragraph (Ground Truth)	59.80	12.80	3.40
RawOCR Paragraph	55.92 (6.49%↓)	5.60 (56.25%↓)	1.00 (70.59%↓)
LLM Corrected Paragraph	54.27 (9.26%↓)	3.40 (73.44%↓)	0.20 (94.12%↓)
RawOCR Corrected Answer	55.21 (7.68%↓)	4.80 (62.50%↓)	0.60 (82.35%↓)

However, in German, substitution errors appear to be less disruptive than in English and French. This can be attributed to the compound word structure in German, where minor substitutions can still preserve some semantic similarity. In contrast, deletions or insertions tend to fragment meaningful lexical units, making them more impactful in German than in other languages.

Summary of findings: Across languages, the results reveal that English and French exhibit similar degradation patterns, with BERTScore progressively decreasing as the OCR error frequency increases. However, in German, the sharpest decline is observed across all error types, particularly for substitutions and deletions. This suggests that OCR noise in German is more detrimental, probably because the older scripts have content in old German where characters such as (long s) are used instead of "s", which can often be misread as "f" or "l".

The results indicate that *substitution errors are the most disruptive in English and French, while German is more affected by deletions* due to its compound word structure. *Insertion errors generally cause moderate degradation*, but severe performance drops occur at high error levels. Overall, *German experiences the highest performance drop*, reinforcing its greater vulnerability to OCR distortions and highlighting the need for effective OCR correction strategies.

5.4 Effectiveness of Error Correction Strategies

Finally, we present an additional study to mitigate the impact of OCR errors on QA through two mitigation approaches, context correction and answer correction of RawOCR text. We want to test simple solutions to reduce the effect of OCR noise in QA while preserving the syntactic and semantic integrity of the input. Such solutions should also be feasible to be deployed in real-world settings with limited computational and financial resources.

We randomly sample 500 questions for each of the three languages. As the context correction approach[11], we correct the RawOCR text first using the Gemma-2 27B model and then use it as context for answering questions. In contrast, in the answer correction approach, we use RawOCR text as context to generate an answer first, and then we correct it afterward.

The context correction approach may improve performance by correcting the entire context before QA. However, it is computationally expensive due to the need to process the entire passage. On the other hand, the Answer correction approach is more lightweight and cost-efficient, as it only involves correcting short answer spans. Nevertheless, it may suffer from context misalignment, especially when the input is both noisy and lacks sufficient contextual information. We use Gemma-2 27B for both strategies as it demonstrated the strongest performance across all languages in our main experiments (Section 5.2). The results of this analysis are summarized for different languages in Table 6, Table 7, and Table 8, respectively. In the tables, the context correction approach is labeled as LLM Corrected Paragraph, while the answer correction approach is labeled as RawOCR Corrected Answer.

The results across English, French, and German (Tables 6, 7, and 8) consistently show that using corrected (ground-truth) paragraphs as context yields the highest QA performance across all metrics. Among the three languages, English shows the greatest benefit from pre-processing RawOCR with LLM correction, with BERTScore and EM metrics improving significantly over using RawOCR directly. In contrast, French and German show minimal or even negative gains from pre-processing, with German showing extremely low EM score with degradation of 94.12% compared to the ground truth. For all three languages, post-processing the answer (i.e., correcting the generated answer after QA) consistently results in the worst performance, suggesting that once the model has reasoned over noisy text, semantic distortions are difficult to recover. Across all languages, the post-processing strategy, where the answer is corrected after being generated from RawOCR input, consistently results in the lowest performance, highlighting that once a model has reasoned over noisy context, semantic errors become difficult to correct.

These findings emphasize the *importance of integrating OCR error correction early in the QA pipeline* to improve the reliability of QA systems, especially when dealing with historical texts or other archival materials of lower quality. However, given the huge collections of digitized content with vastly varying levels of OCR quality that the current memory institutions (archives, libraries, museums, etc.) held, the correction cost and effort would be enormous. It is also difficult to correct the queried texts at inference time as this would also introduce computational cost and latency in online systems. Therefore, more robust QA systems that are aware of OCR errors and capable of predicting correct answers based on contextual information are required.

6 Conclusion

In this paper, we conducted a comprehensive evaluation of QA systems under the influence of OCR-induced noise using multilingual historical texts. To facilitate this analysis, we introduced `MultiOCR-QA`, the first multilingual QA dataset derived from OCR-processed historical documents in English, French, and German. The dataset includes both RawOCR (noisy) and CorrectedOCR (clean) text, enabling controlled and comparative assessments of model robustness under real-world digitization errors. By leveraging both CorrectedOCR and RawOCR we systematically analyzed how different types of OCR errors—insertions, deletions, and substitutions affect the performance of LLMs in QA.

Our experiments reveal that OCR noise leads to substantial degradation in QA performance across all languages and metrics, with particularly severe impacts in settings with high error rates. While larger LLMs such as Gemma-2 27B and Qwen-2.5 72B show relatively better robustness, even these models experience significant accuracy drops when processing noisy text. Moreover, pre- and post-processing strategies using LLMs to correct OCR errors offer only limited benefits and can, in some cases, degrade performance even further. These findings highlight the limitations of current QA systems in handling real-world noisy text and emphasize the need for evaluation frameworks and model designs that can handle OCR distortions, especially in historical and multilingual contexts.

Use cases of `MultiOCR-QA`: The `MultiOCR-QA` dataset offers a unique resource to advance research on OCR-aware QA and studying QA on noisy OCR text, making it useful in several ways. It can be used to train LLMs to improve error correction capabilities and enhancing robustness against OCR inaccuracies while preserving the archaic language structure. It can also be used to expand LLMs' multilingual processing abilities by training on OCR text in multiple languages, enhancing performance in languages beyond English.

Limitations and Future work: While `MultiOCR-QA` includes English, French, and German, it does not encompass low-resource languages or scripts such as Latin, Finnish, and others. Future research should incorporate low-resourced languages to improve generalization and greater applicability across diverse languages. Additionally, methodologies that not only remove OCR errors but also preserve the original structure of documents could be applied.

Acknowledgments

The computational results presented here have been achieved (in part) using the LEO HPC infrastructure of the University of Innsbruck. This Project was Co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 — AI4DH. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

[11] The prompt used for context correction. **System Prompt:** *You are an expert at understanding the historical texts and correcting OCR errors.* **User Prompt:** *You are provided with a historical English text containing spelling mistakes. Correct only the spelling mistakes and present the corrected text in a single paragraph. If you cannot correct the mistakes, reply with "Not able to correct." Historical Text : {context}*

Usage of Generative AI

This work involved the use of Generative AI (GenAI) models at several stages of the research process. The MultiOCR-QA dataset has been automatically generated using an instruction fine-tuned large language model (LLaMA 3.1-70b) for English, French, and German. The models were explicitly used for answer span extraction and question generation based on historical CorrectedOCR text. Various generative models, including LLaMA, Qwen, Mixtral, and Gemma, were used in the downstream evaluation of QA performance across corrected and noisy OCR text. Additionally, GenAI tools such as ChatGPT and Gemini were occasionally used to assist with resolving coding bugs or syntax errors in scripts or LaTeX syntax.

All conceptual work, dataset engineering, code design, and writing remain the sole intellectual contributions of the authors. GenAI tools were used in a supportive capacity, similar to that of debugging assistants or grammar checkers, without contributing novel research content.

References

[1] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: towards a better access to information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (Toronto, Ontario, Canada) *(JCDL '17)*. IEEE Press, 249–252.

[2] W Bruce Croft, SM Harding, Kazem Taghva, and Julie Borsack. 1994. An evaluation of information retrieval accuracy with simulated OCR output. In *Symposium on Document Analysis and Information Retrieval*. 115–126.

[3] Lucas Lima de Oliveira, Danny Suarez Vargas, Antônio Marcelo Azevedo Alexandre, Fábio Corrêa Cordeiro, Diogo da Silva Magalhães Gomes, Max de Castro Rodrigues, Regis Kruel Romeu, and Viviane Pereira Moreira. 2023. Evaluating and mitigating the impact of OCR errors on information retrieval. *Int. J. Digit. Libr.* 24, 1 (Jan. 2023), 45–62. doi:10.1007/s00799-023-00345-6

[4] Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French Question Answering Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1193–1208. doi:10.18653/v1/2020.findings-emnlp.107

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[6] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *Comput. Surveys* 56, 2 (2023), 1–47.

[7] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*. Springer, 288–310.

[8] Faisal Farooq[1], SUNY CEDAR, and Yaser Al-Onaizan. 2005. Effect of degraded input on statistical machine translation. In *Proceedings 2005 Symposium on Document Image Understanding Technology*. UMD, 103.

[9] Edward Giamphy, Kevin Sanchis, Gohar Dashyan, Jean-Loup Guillaume, Ahmed Hamdi, Lilian Sanselme, and Antoine Doucet. 2023. A Quantitative Analysis of Noise Impact on Document Ranking. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 4612–4618.

[10] Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named Entity Recognition for Digitised Historical Texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Marrakech, Morocco. https://aclanthology.org/L08-1253/

[11] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Assessing and minimizing the impact of OCR quality on named entity recognition. In *Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings 24*. Springer, 87–101.

[12] Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2022. In-Depth Analysis of the Impact of OCR Errors on Named Entity Recognition and Linking. *Journal of Natural Language Processing* (03 2022), 24. doi:10.1017/S1351324922000110

[13] Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2023. In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Natural Language Engineering* 29, 2 (2023), 425–448.

[14] Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities* 34, 4 (2019), 825–843.

[15] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[16] Hongyan Jing, Daniel Lopresti, and Chilin Shih. 2003. Summarizing noisy documents. In *Proceedings of the Symposium on Document Image Understanding Technology*. 111–119.

[17] Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda (Eds.). Association for Computational Linguistics, Melbourne, Australia, 74–83. doi:10.18653/v1/W18-2709

[18] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28186–28194.

[19] Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep Learning for Period Classification of Historical Hebrew Texts. *J. Data Min. Digit. Humanit.* 2020 (2020). https://api.semanticscholar.org/CorpusID:225718181

[20] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. 2019. Impact of OCR quality on named entity linking. In *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings 21*. Springer, 102–115.

[21] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press.

[22] Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, Adam Fisch, Alon Talmor, Danqi Chen, Eunsol Choi, Minjoon Seo, Patrick Lewis, Robin Jia, and Sewon Min (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 42–50. doi:10.18653/v1/2021.mrqa-1.4

[23] Stephen Mutuvi, Antoine Doucet, Moses Odeo, and Adam Jatowt. 2018. Evaluating the impact of OCR errors on topic modeling. In *International Conference on Asian Digital Libraries*. Springer, 3–14.

[24] Michael Piotrowski. 2012. *Natural language processing for historical texts*. Vol. 17. Morgan & Claypool Publishers.

[25] Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2038–2048. doi:10.1145/3626772.3657891

[26] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2383–2392. doi:10.18653/v1/D16-1264

[27] 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)*. 1588–1593. doi:10.1109/ICDAR.2019.00255

[28] Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. 2018. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. *J. Lang. Technol. Comput. Linguistics* 33, 1 (2018), 97–114. https://jlcl.org/content/2-allissues/1-heft1-2018/jlcl_2018-1_5.pdf

[29] Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534* (2020).

[30] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[31] MELISSA M TERRAS. 2011. 1. THE RISE OF DIGITIZATION. *Digitisation Perspectives* (2011), 3.

[32] Myriam C Traub, Jacco Van Ossenbruggen, and Lynda Hardman. 2015. Impact analysis of OCR quality on research tasks in digital archives. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory*

and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19. Springer, 252–263.

[33] Daniel Alexander van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *International Conference on Agents and Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:215756646

[34] Oxana Vitman, Yevhen Kostiuk, Paul Plachinda, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Evaluating the Impact of OCR Quality on Short Texts Classification Task. In *Mexican International Conference on Artificial Intelligence*. Springer, 163–177.

[35] Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, et al. 2023. Are large language models really robust to word-level perturbations? *arXiv preprint arXiv:2309.11166* (2023).

[36] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[37] Tze-I Yang, Andrew Torget, and Rada Mihalcea. 2011. Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Kalliopi Zervanou and Piroska Lendvai (Eds.). Association for Computational Linguistics, Portland, OR, USA, 96–104. https://aclanthology.org/W11-1513/

[38] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 2088, 15 pages.

[39] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792* (2023).

[40] Elaine Zosa, Stephen Mutuvi, Mark Granroth-Wilding, and Antoine Doucet. 2021. Evaluating the robustness of embedding-based topic models to OCR noise. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23*. Springer, 392–400.

[41] Guowei Zu, Mayo Murata, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura. 2004. The impact of OCR accuracy on automatic text classification. In *Content Computing: Advanced Workshop on Content Computing, AWCC 2004, ZhenJiang, JiangSu, China, November 15-17, 2004. Proceedings*. Springer, 403–409.