

Digitizing Nepal’s Written Heritage: A Comprehensive HTR Pipeline for Old Nepali Manuscripts

Anjali Sarawgi^{1*} Esteban Garces Arias^{1,3} Christof Zotter²

¹Department of Statistics, LMU Munich, ²Heidelberg Academy of Sciences and Humanities

³Munich Center for Machine Learning (MCML)

*Corresponding author: Anjali.Sarawgi@campus.lmu.de

Abstract

This paper presents the first end-to-end pipeline for Handwritten Text Recognition (HTR) for Old Nepali, a historically significant but low-resource language. We adopt a line-level transcription approach and systematically explore encoder-decoder architectures and data-centric techniques to improve recognition accuracy. Our best model achieves a Character Error Rate (CER) of 4.9%. In addition, we implement and evaluate decoding strategies and analyze token-level confusions to better understand model behavior and error patterns. Although the evaluation dataset is confidential, we release our training code, model configurations, and evaluation scripts to support further research on HTR for low-resource historical scripts.

1 Introduction

Handwritten Text Recognition (HTR) remains a challenging task—particularly for historical and low-resource languages such as Old Nepali. Historical documents pose numerous challenges for offline handwriting recognition systems, especially during the segmentation and labeling stages (Wigington et al., 2018; Chammas et al., 2018). Nepal, particularly the Kathmandu Valley, has a rich manuscript culture that features diverse scripts for various languages, including Newari (also referred to as Nepal Bhasha), Sanskrit, and Maithili. For the text recognition of Pracalit Lipi, the most common script in the Valley for centuries, we refer to Nakarmi et al. (2024). The historical manuscripts used in this study are written in Old Nepali, the language of the Gorkhals who conquered the Kathmandu Valley kingdoms in the late 18th century. The new government used the Devanagari script, which is still dominant in modern Nepal and many parts of India.

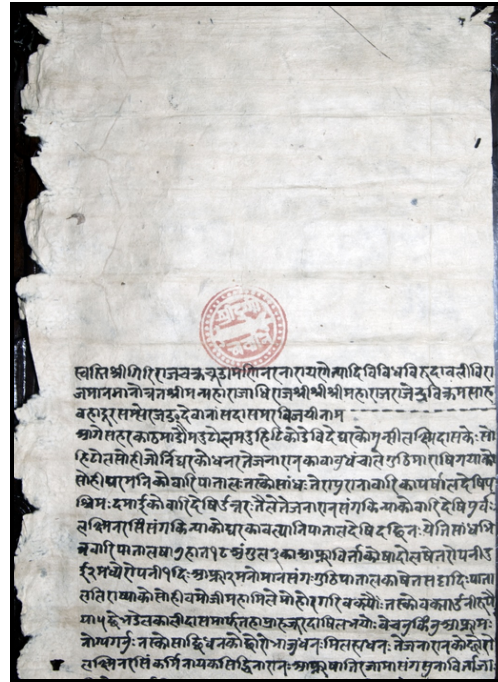


Figure 1: Sample manuscript containing Old Nepali in Devanagari script. This image is sourced and cropped from the Documenta Nepalica (Śāha, 1832), courtesy of Manik Bajracharya.

However, the historical manuscripts considered in this study present several challenges, including diverse handwriting styles, degraded document quality, intricate conjunct forms, and limited annotated data (Nockels et al., 2024). These constraints necessitate approaches that leverage transfer learning, data augmentation, and carefully designed model architectures (Garces Arias et al., 2023). In this paper, we address these challenges by conducting a comprehensive exploration of modern HTR techniques for Old Nepali manuscripts. We investigate transfer learning strategies to enable effective learning in low-resource settings, implementing a three-stage approach that adapts models from high-resource settings to target-domain data.

Our architectural exploration includes transformer-based approaches such as TrOCR (Li et al., 2023), as well as combinations of vision encoders (including Swin Transformer (Liu et al., 2021)) with language decoders (BERT-style (Devlin et al., 2019) and GPT-2 (Radford et al., 2019)). To effectively capture the intricacies of Old Nepali scripts, we compare byte-level tokenization with traditional approaches to determine the optimal granularity for this complex writing system. We further enhance model robustness through data augmentation strategies that preserve script integrity while improving generalization (Rassul et al., 2025). Finally, we systematically compare decoding methods—including deterministic methods (greedy, beam search (Freitag and Al-Onaizan, 2017), contrastive search (Su et al., 2022)) and stochastic methods (sampling with temperature (Ackley et al., 1985), top- k (Fan et al., 2018), and top- p sampling (Holtzman et al., 2019))—to optimize recognition accuracy.

Contributions: (i) We develop a comprehensive HTR pipeline that efficiently performs recognition on Old Nepali texts, demonstrating superior performance compared to both open-source and proprietary alternatives in this low-resource setting. (ii) We present a detailed analysis of preprocessing and data augmentation methods specifically tailored for historical scripts. (iii) We demonstrate the impact of domain adaptation through pretraining on datasets derived from publicly available sources, including the Heidelberg open research data (Merkel-Hilf, 2022). (iv) We conduct a thorough investigation of tokenization and decoding methods to capture the unique structural properties of historical Nepali texts, revealing insights about optimal representation strategies for complex scripts. (v) We make our code, models, and analysis tools publicly available at <https://github.com/anjalisarawgi/nepOCR/>.

2 Related Work

HTR has evolved from traditional pattern recognition methods to deep learning approaches. Early systems were based on RNNs and the CTC framework developed by Graves and Schmidhuber (2008), while later works combined CNNs with RNNs and CTC decoding (Shi et al., 2017).

The advent of Tesseract OCR (Smith, 2007) provided a widely used baseline, albeit with limited accuracy for historical scripts. Recent advances in HTR have been driven by transformer-based

architectures and deep learning approaches (AIK-endi et al., 2024). TrOCR (Li et al., 2023), which leverages the Transformer architecture for both image understanding and wordpiece-level text generation, has shown significant improvements over traditional CNN-RNN approaches (Garces Arias et al., 2023). However, applying these state-of-the-art methods to low-resource historical scripts requires careful adaptation. Handwritten text recognition in low-resource scenarios, such as manuscripts with rare alphabets, is a challenging problem (Souibgui et al., 2022).

In addition to model architectures, recent research has focused on data augmentation for low-resource scripts (Gupta et al., 2016) and the use of synthetic data to overcome the scarcity of annotated manuscripts. Prior studies (Wigington et al., 2017; de Sousa Neto et al., 2024) have also shown the value of data augmentation for HTR tasks. Tokenization methods, a critical component of natural language processing, have been extensively explored in the context of subword regularization (Kudo, 2018) and have shown promise for adapting language models to complex scripts. Recent studies have also examined the interplay between vision encoders and language decoders. For example, combinations involving Swin Transformers with BERT or GPT-2 have demonstrated promising results in text recognition tasks (Garces Arias et al., 2023; Koch et al., 2023; Pavlopoulos et al., 2024). Furthermore, novel decoding strategies such as contrastive search, adaptive contrastive search (Garces Arias et al., 2024), GUARD (Ding et al., 2025), and Min- k (Ding et al., 2026) have been proposed to overcome the limitations of beam search across various NLP tasks.

3 Data Preparation

3.1 Handwritten scripts

The majority of the manuscripts analyzed in this study originate from the collection of the Nepal German Manuscript Preservation Project (NGMPP).¹ The NGMPP undertook extensive microfilming of manuscripts and related documentary materials from a range of Nepalese repositories, including the National Archives in Kathmandu, as well as private collections. The selected examples date to the late eighteenth and nineteenth centuries, with many consisting of royal edicts bearing the red

¹For further details, see Documenta Nepalica: <https://nepalica.hadw-bw.de/nepal/>.

seal of the Shah kings. They typically have a consistent visual and structural layout (as illustrated in Figure 1) and follow the diplomatic conventions of the period. They begin with an invocation to a deity, followed by a blank space where the royal seal is imprinted. The main text begins with the standard Sanskrit panegyric of the king, names the addressee, discusses the subject matter of the document, and ends with the *eschatocol* containing the date and a blessing (Cubelic et al., 2018; Pant and Pierce, 1989). The selection also includes well-written petitions to the king that share some diplomatic elements but typically have a simpler structure. Most of these documents consist of one or two pages, providing a limited number of training samples. To improve the dataset, we also included larger samples from two chronicle manuscripts that share a similar context and writing style.

As with South Asian manuscript cultures in general, *scriptio continua* is the norm here. Only a few more recent documents include spaces to separate words. Hand-drawn bullets or other punctuation symbols, such as dandas, may appear at the end of a word, sentence, or clause. However, these symbols are used inconsistently, and bullets can also appear within words. As a result, detecting word and sentence boundaries remains challenging.

Devanagari is a rich, left-to-right abugida script and includes many complex conjunct characters (cf. Figures 17 and 18). The form used in the documents differs slightly from present-day Devanagari. In addition, variations in handwriting, scribal styles, and even pen types introduce additional inconsistencies, making text recognition more challenging. Although the texts may include passages in other languages (such as the Sanskrit panegyrics), the main language is Old Nepali. Morphologically, this stage of language development is quite close to modern Nepali but differs, especially in orthography, which is highly variable (Riccardi, 1971). Due to a lack of standardization, scribes often spelled the same word differently, even within a single document. Another challenge is the abundance of now-obsolete technical terms. The transcribed dataset used in this study consists of 155 manuscript images. The manuscripts are written in Devanagari, and an average page contains 1,198 characters across 20 lines. The lines are relatively long, averaging 60 characters. The average dimensions of the manuscript images are 3091×3487 pixels at 328 dpi, and a detailed character set is provided in Appendix G.

3.2 Image preprocessing

We preprocess the Old Nepali dataset using line-level segmentation with Kraken’s (Kiessling, 2019) polygon-based segmentation method to detect and isolate individual text lines in the manuscript images. Each detected line is then cropped and used as an independent training sample, allowing the model to focus solely on text recognition. From a total of 155 manuscripts, we extracted 3,100 lines. The longest sequence contains 162 characters, and the shortest line consists of only 1 character. The standard deviation of line lengths is 27.65 characters, indicating high variation. The images are typically very wide and short, with an average size of 1593×133 pixels (width \times height). An example of a cropped line image after preprocessing is shown in Figure 2, and detailed statistics are provided in Appendix F.



Figure 2: Sample line image after preprocessing.

3.3 Three-stage training dataset preparation

Our training pipeline comprises three stages that adapt the model to Old Nepali manuscripts (see Appendix H). This setup is important for two reasons: (1) the limited availability of labeled handwritten Devanagari data and (2) the need for a model that understands Devanagari script and its visual and stylistic variations. In the **first stage**, the model is trained on a large synthetic dataset to learn general Devanagari patterns. The **second stage** uses a printed Devanagari dataset to bridge the gap between synthetic and real data. Finally, in the **third stage**, the model is fine-tuned on the Old Nepali manuscript dataset. Here, the first and second stages function as pretraining to enable the model to learn general context and visual patterns, and the third stage performs final fine-tuning on the manuscript data.

First stage We train on a synthetic dataset constructed using text extracted from historical Nepali textbooks available through the Internet Archive (Internet Archive, 2025). The books from which the text corpus is extracted are listed in Table 16, in Appendix K. These texts are rendered as line-level grayscale images using the Python Imaging Library (PIL), with a variety of Devanagari fonts. We generate 105,000 images and simulate script-

level degradation by applying several noise and distortion effects. Specifically, we use a total of 11 fonts and 10 types of noise and distortion effects (e.g., geometric distortions, blurring, contrast or brightness variations), and the choice of font type and augmentation method for each image is randomized. A detailed overview is provided in Table 14 and Table 15 in Appendix K.

This synthetic dataset is prepared to mirror the real dataset. All images are at the line level, as we render individual lines of a textbook. From this synthetic corpus, we use 100,000 images for training and set aside 2,500 images each for validation and testing. This stage serves as an initial training phase for the decoder, helping it learn domain-relevant linguistic patterns. A sample from the synthetic dataset used in the first stage is shown in Figure 3.

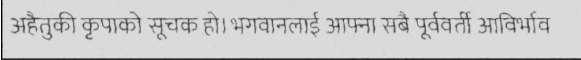


Figure 3: Sample data for the first stage.

Second stage We train on a printed Devanagari dataset sourced from heiDATA (Merkel-Hilf, 2022), the open research data repository of Heidelberg University. As with the Old Nepali manuscripts, we apply preprocessing and line-level segmentation to scanned pages, yielding 5,139 line images from 247 full-page scans. The images are converted to grayscale to match the visual appearance of the Old Nepali manuscript data. We use 80% of these line images for training, and allocate 10% each for validation and test sets. While this dataset contains printed rather than handwritten script, it includes realistic noise from scanned script documents, making it useful as a transfer stage to bridge the gap between synthetic and handwritten data. A sample from the printed Devanagari dataset used in the second stage is shown in Figure 4.



Figure 4: Sample data for the second stage.

Third stage We fine-tune on the target Old Nepali manuscript dataset of 3,100 line-level images from 155 historical manuscripts (80/10/10 train/validation/test split). This final stage adapts the model to the specific handwriting styles, orthographic variations, and degradation patterns unique to these historical documents. Table 1 summarizes

the data split across all training stages. For a fair comparison, we split the dataset once with a fixed seed and utilize the same partitions.

	Train	Eval	Test
First Stage: Pre-training (Synthetic Devanagari images)	100,000	2500	2500
Second Stage: Transfer Learning (Printed Nagari scripts)	4,111	514	514
Third Stage: Final Model Training (Handwritten Old Nepali images)	2,480	310	310
Total	106,591	3,324	3,324

Table 1: Data configuration for the three-stage training.

4 Experimental Setup

This section outlines the model configurations and data refinement strategies used in our experiments. We first compare TrOCR encoder variants, language decoders, and tokenizer types. We then introduce a set of data-centric modifications designed to boost model performance in the low-resource setting of Old Nepali handwritten text.

4.1 Model selection

We use two TrOCR variants, trocr-base-handwritten and trocr-large-handwritten, which are ViT-based encoders pretrained on printed and handwritten datasets. A Swin Transformer encoder is also evaluated as an alternative architecture. For decoders, we explore two types of language models, GPT-2 and BERT, and evaluate different tokenization methods to assess their effectiveness for Devanagari scripts. This is motivated by the need for Devanagari-aware decoding and by the question of whether script-specific tokenization improves performance. We focus on transformer-based models, as prior work, e.g., Garces Arias et al. (2023), shows that they outperform CNN-RNN architectures on historical HTR tasks.

The combinations of encoder, decoder, and tokenizer result in a total of $3 \times 2 \times 2 = 12$ model configurations. Each configuration is trained using our three-stage setup: 6 epochs for pretraining, 10 for transfer learning, and 20 for final fine-tuning, totaling $12 \times 3 = 36$ runs. All stages use a fixed learning rate of $3e-5$ with the AdamW optimizer, and a batch size of 8 (see Table 2 for more details). Table 7 in Appendix C presents the results for all model configurations. Model selection is based on validation performance at the final fine-tuning stage, where the model with the lowest Character Error Rate (CER) is selected.

Parameter	First Stage	Second Stage	Third Stage
Seed	42	42	42
Optimizer	AdamW	AdamW	AdamW
Epochs	6	10	20
Batch size (Train/Val/Test)	8	8	8
Learning rate	3e-5	3e-5	3e-5
Weight decay	0.01	0.01	0.01
Warmup steps	500	500	500

Table 2: Training hyperparameters for all three learning stages.

4.2 Data-centric enhancements

After selecting the best-performing model architecture, we conduct a series of controlled experiments to measure the impact of preprocessing and dataset refinement on the final model performance. Since our labeled dataset is very small, we focus on strategies that help the model maximize its learning on each training sample. Specifically, we apply data augmentations to introduce a wider range of visual variations during training, thereby expanding dataset diversity and improving model generalization. We also perform light label cleaning to reduce noise and inconsistencies in the ground truth that might confuse the model and distract it from learning meaningful character-level patterns.

Binarization We evaluate the impact of binary thresholding as a preprocessing step for line images. Contrary to our hypothesis that binarization would improve model performance, the results indicate otherwise, with binarization slightly worsening performance by 1% CER (see Appendix B for details). Thus, we decided to use non-binarized images for training.

Transcription normalization To ensure consistency and stable training, we apply a series of stylistic, punctuation, and Unicode normalization steps to the transcriptions. This is particularly important in Old Nepali HTR, where even minor transcription inconsistencies can introduce ambiguity and degrade the model’s performance. For example, characters such as U+0310 and U+0901 both represent the Devanagari chandrabinu but may appear inconsistently across samples, introducing noise and inconsistencies during training. Given the limited size of our dataset, these inconsistencies introduce unnecessary variation that can hinder the model’s learning. More details on the normalization steps are provided in Appendix I.

Data Augmentation To enhance model robustness against variability in handwriting, script style, and scan quality, and to expand the effective size

of the training set through synthetic samples, we apply a wide range of image-level augmentations. These augmentations are moderate and realistic, aiming to mimic the distortions and degradations commonly found in handwritten historical scripts, without damaging the underlying script structure.

We group our augmentations into three categories. The first group includes *shape and angle distortions*, such as small-angle rotations, axis shifts, horizontal and vertical stretching, and image tilts and skews, to simulate images acquired at oblique angles. These simulate camera angle, scanning inconsistencies, or physical wrapping of the document. The second group comprises quality-degradation augmentations, including Gaussian blur and noise, blurring in one direction, salt-and-pepper noise, JPEG compression artifacts, and jitter to introduce pixel-level noise. These transformations simulate common sources of noise arising from scanning hardware, paper texture, or manuscript age. The third group focuses on character-level distortions and is comparatively stronger than the other two groups. It includes smudging random parts of the text with small blurring boxes, elastic warping (to mirror crumpled pages), subtle text curvature (as from slightly bent documents), and mild morphological operations such as erosion, dilation, and edge sharpening. These help the model handle character-level degradations frequently observed in noisy or damaged manuscripts.

We use 20 augmentation variations and apply augmentation intensities of 2 \times , 4 \times , 8 \times , 12 \times , and 16 \times . A detailed list of augmentation functions is provided in Appendix J.

4.3 Model Architecture

Tokenizer We train two types of subword tokenizers using the Hugging Face tokenizers library (Moi and Patry, 2023): CharBPETokenizer and ByteLevelBPETokenizer. Both implement the Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2016), which iteratively merges frequent token pairs until a target vocabulary size is reached. For our experiments, we fix the vocabulary size to 500 to balance the trade-off between coverage of textual variation with token frequency in training, which is particularly beneficial in low-resource settings (Garces Arias et al., 2023). To train the tokenizers, we construct a combined corpus comprising lines from all three stages of our HTR pipeline.

Image encoder We adopt the trocr-base-handwritten model as our primary image encoder. It uses a 12-layer Vision Transformer (ViT), and processes images resized to 384×384 pixels. The encoder splits images into patches, converts them to vectors, and processes them through Transformer layers to learn visual features for text recognition. Although this model was not explicitly trained on the Devanagari script, its pretraining on diverse handwritten and printed scripts enables it to generalize reasonably well to handwritten Devanagari forms, especially given the lack of a large labeled dataset of handwritten Devanagari. To test architectural variants, we also evaluate a Swin Transformer encoder (swin-base-patch4-window7-224-in22k), which underperformed due to limited handwritten Devanagari data (see Appendix D). In our work, the TrOCR image encoder remained the better choice.

Text decoder We evaluate two decoder architectures for text generation: a BERT-based model (114.8M parameters) and a GPT-2-based model (114.6M parameters). Both models are trained from scratch on the synthetic dataset in the first stage (see Table 1). To better adapt to our data, we use custom script-aware tokenizers designed for Devanagari text. The decoders follow standard base configurations (12 layers, 768 hidden size, 12 attention heads), with a language modeling head added for token prediction. The only modification is the vocabulary size, which is set to match our custom tokenizer. They are also paired with the same visual encoder, which allows direct comparison of decoder performance for HTR.

Performance metrics We evaluate model performance using CER, weighted CER (CER(w)), and exact match accuracy (ACC), i.e., the percentage of predicted lines that exactly match the text. In this work, we use CER as the primary metric for all model design choices. The CER is computed as:

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

where S , D , I denote substitutions, deletions, and insertions respectively, and C denotes correct characters. Additionally, to account for the varying length of the labels, which range from 1 to 162, we

utilize the weighted CER:

$$weighted\ CER = \frac{\sum_{i=1}^n l_i \cdot CER_i}{\sum_{i=1}^n l_i} \quad (2)$$

where l_i is the number of characters of label i , and CER_i is the CER for example i , $i = 1, \dots, n$.

During evaluation, we remove zero-width Unicode characters (`\u200B`, `\u200C`, and `\u200D`) from both predictions and ground truth labels before computing CER. These characters are used in Devanagari to control conjunct forms and are present in the training data. However, since they are often added for rendering, we exclude them to focus on core transcription accuracy.

Decoding strategies We test five decoding methods: beam search, contrastive search, temperature sampling, top- k , and top- p sampling. For each strategy, we evaluate a range of hyperparameters to study their impact on HTR performance. A detailed overview is provided in Appendix L.

5 Results

5.1 Model selection

Table 7 in Appendix C presents the results of our comparative experiments across different decoder architectures and tokenization schemes. Overall, the choice of decoder architecture has only a small effect on final model performance. Interestingly, even the tokenizer appears relatively robust to architectural changes. Among all configurations, the BERT decoder paired with the byte-level BPE achieves the best performance, reaching a CER of 0.082 after all three training stages. This suggests that architectural tuning yields only limited gains, with BERT and byte-BPE achieving the most favorable CER.

5.2 Performance gains through label cleaning and augmentation

We evaluated the impact of label normalization and image-level augmentations on model performance. As shown in Table 3, label cleaning alone reduced CER from 0.089 to 0.084. The most substantial gains were achieved through data augmentation, which reduced CER to 0.056 with $8 \times$ augmentation. Increasing it to $12 \times$ and $16 \times$ did not yield further improvements, indicating a performance plateau. These results confirm the importance of both label quality and visual diversity for improved performance in low-resource OCR tasks.

Ablation Step	Samples	CER	CER(w)	ACC
Original Dataset	2,480	0.089	0.090	22.9%
+ Normalization	2,480	0.084	0.084	21.6%
+ Augmentation (2×)	7,440	0.067	0.068	26.7%
+ Augmentation (4×)	12,400	0.060	0.061	27.1%
+ Augmentation (8×)	22,320	0.056	0.057	29.4%
+ Augmentation (12×)	32,240	0.056	0.056	30.0%
+ Augmentation (16×)	42,160	0.056	0.056	29.7%

Table 3: Ablation study showing the impact of normalization and data augmentation on model performance. The best results are shown in **bold**.

Evaluating impact of encoder variants We compared `trocr-base-handwritten` and `trocr-large-handwritten` encoders to assess the effect of model size. As shown in Table 4, the larger variant achieved improved results by reducing CER from 0.056 to 0.049 with a corresponding weighted CER of 0.048.

Encoder Variant	CER	CER(w)	ACC
<code>trocr-base-handwritten</code>	0.056	0.057	29.4%
<code>trocr-large-handwritten</code>	0.049	0.048	33.5%

Table 4: Performance comparison of TrOCR encoder variants (base and large) under identical training conditions. The best results are shown in **bold**.

5.3 Evaluating decoding methods

We observe that different decoding strategies and hyperparameter settings have minimal effect on the overall performance. For instance, all beam sizes (1, 5, 10, 20) yield an identical weighted CER of 0.0483. Contrastive search with varying k and α , as well as sampling-based methods with different τ , k , and p values, consistently result in $\text{CER} \approx 0.0488$. These negligible differences (Appendix L) suggest that the decoding strategy has minimal impact on the achieved HTR performance.

5.4 Benchmarking against HTR & OCR tools

We benchmark our pipeline against two baselines: (1) a full TrOCR model (including both encoder and decoder), and (2) Google Cloud Vision OCR. Since TrOCR is not pretrained on Devanagari scripts, it produces non-Devanagari outputs. To ensure a fair comparison, we fine-tuned TrOCR on our dataset using the same training configuration as our approach. The fine-tuned TrOCR baseline achieves a CER of 0.096. Additionally, Google Cloud Vision OCR performs poorly on our dataset, failing to capture conjuncts, diacritics, and punctuation (see Appendix M for a detailed overview).

Furthermore, the default TrOCR decoder, when fine-tuned on the normalized and 8x-augmented dataset, achieves a CER of 0.059. In comparison, our decoder achieves a slightly lower CER of 0.056, with notably fewer parameters and $1.4\times$ faster evaluation, indicating better efficiency without compromising accuracy (see Appendix E).

5.5 Error analysis

To assess the strengths and weaknesses of our model, we conduct an error analysis focusing on token confusions and label lengths (see Appendix O for visualizations). We construct confusion matrices by comparing model predictions on the test set with the corresponding ground truth labels, thereby identifying frequently misclassified characters. Inspection of the resulting heatmap shows that many errors are structured rather than random, following systematic patterns of handwriting variation. For example, य (ya) is often confused with प (pa), and त (ta) with न (na). This is likely because these character pairs exhibit high visual similarity in handwritten form, with the degree of similarity varying across handwriting styles. This suggests that the model is not hallucinating but making plausible mistakes due to visual ambiguity. Such structured errors can also be leveraged for post-correction methods. Furthermore, we examined which individual characters the model makes the most errors on. This analysis revealed that a small subset of 10 characters (out of 80 total) accounts for over half of all errors, i.e., 55.9% (Figure 14 and Table 19). This highlights the specific characters the model consistently struggles with and can be prioritized in future work. The overall distribution of CER across individual lines in the test set is presented in Figure 11. We observe a strongly skewed distribution, with most lines exhibiting low CERs in the range 0–0.05. This indicates that, for most samples, the model predictions are highly accurate and require minimal or no correction. Additionally, we analyze how model performance varies with line length (see Figure 13). The weighted CER remains relatively low and stable across short and medium-length lines. However, the error rates increase sharply for lines longer than 120 characters. This spike is likely due to both the increased difficulty of longer sequences and the limited number of such samples, with only 26 out of 3,100 total lines exceeding 120 characters. We further analyzed the error patterns for longer lines and found two common patterns: (i) missing tokens at multiple positions and (ii) short

segments (3–7 characters) with consecutive incorrect predictions. To investigate this, we segmented these lines into two parts and re-evaluated them using our model. Here, we observed a substantial drop in errors; for example, in one case, errors decreased from 23 to 4, and in another, from 30 to 2. This finding suggests that the model has difficulty handling longer sequences, and splitting them into shorter parts can improve model performance. Finally, the script-aware decoder enables us to explore model uncertainty by comparing the relative probabilities of the top predictions. This helps us flag about 27% of all errors, and over half of these can be recovered from the top-3 predictions. This shows potential for future work in post-correction and human-in-the-loop review (see Appendix P).

5.6 Inspecting domain shift

Pretraining on the printed Nagari script proved highly effective for our HTR task. Once fine-tuned, the model generalized well to handwritten data, despite the small size of the labeled dataset. This transfer learning approach was crucial for handling the low-resource setting. Figures 5 and 6 show the visual contrast between the printed and handwritten inputs the model was trained on, demonstrating how it adapted from clean printed forms to variable handwritten forms.

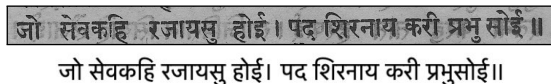


Figure 5: Printed Nagari script sample (top) and model’s predicted transcription (bottom). The prediction is an exact match.



Figure 6: Old Nepali script sample (top) and model’s predicted transcription (bottom). Mistakes are highlighted in red.

We also evaluate the model at each training stage on the final test set to measure the contribution of each stage to the final model performance. The results in Table 18 in Appendix N show that all three stages contribute positively to the model’s final performance.

6 Old Nepali Handwritten Text Recognition Application

To demonstrate the practical applications of our approach, we have developed a web-based application that enables users to process images of Old Nepali manuscripts. This interactive tool allows researchers, historians, and linguists to upload photographs or scans containing Old Nepali handwritten text and receive automated analysis results. The application performs two primary functions: (1) automatic line detection, which segments the manuscript into individual text lines, and (2) text recognition, which converts the handwritten content into machine-readable text. The system provides users with a visual overlay that displays the detected line boundaries alongside the recognized text, facilitating verification. The application interface and sample processing results are illustrated in Appendix S. The web application is available at <https://nepocr.streamlit.app/>.

7 Discussion and Outlook

Our experiments demonstrate advances in handwritten text recognition for low-resource historical scripts. The CER decreased by 49%, from 9.6% with the fine-tuned TrOCR baseline to 4.9% with our final model. This improvement resulted from systematic enhancements to the HTR pipeline, as detailed in Figure 16 (Appendix Q). The most substantial gains were achieved through three key innovations: script-aware decoder architecture, transcription normalization, and data augmentation combined with an enhanced vision encoder. These results confirm that modern transformer architectures can effectively adapt to non-Latin scripts when properly configured. Our three-stage training approach proved particularly valuable, with synthetic data bridging the gap between pretrained models and limited historical data. Interestingly, the choice of decoder architecture (BERT-style vs. GPT-2) and tokenization scheme had minimal impact on performance. This finding suggests that in low-resource settings, data quality and augmentation strategies matter more than model capacity. Indeed, our data-centric approaches—label normalization and data augmentation—produced the largest performance improvements. Error analysis provided additional insights into model behavior. Most character confusions follow predictable patterns based on visual similarity (e.g., य vs. ष), indicating that the model learns meaningful visual rep-

representations rather than memorizing patterns. However, performance degrades on lines exceeding 120 characters, likely due to insufficient long-sequence examples in our training data. Several promising directions for future research emerge. First, collaborating with digital archives to expand the training corpus could help address current limitations in long-sequence data. Second, post-correction methods that exploit structured error patterns may further reduce CER, though this remains challenging because the dataset lacks explicit word boundaries. Third, applying our approach to other historical scripts would validate its generalizability.

8 Conclusion

We present a comprehensive pipeline for recognizing handwritten Old Nepali manuscripts, achieving a CER of 4.9% through the systematic application of domain adaptation, data augmentation, and preprocessing. Our three-stage training approach helps bridge the gap between modern OCR models trained on Latin scripts and the specific challenges of historical Devanagari manuscripts, enabling effective learning in a low-resource setting. The experimental results suggest that data-centric improvements—particularly augmentation and label normalization—can contribute more to performance gains than architectural variations in this context, underscoring the importance of carefully curated training data. The model exhibits robust behavior across a broad range of handwriting styles and short-to-medium line lengths, with errors following interpretable patterns based on visual character similarity. These observations indicate that the model captures meaningful visual and structural features of the script, although some limitations remain, particularly for longer sequences and rare character combinations. This highlights the importance of expanding both the size and diversity of annotated datasets, as well as exploring complementary strategies such as post-correction or sequence-level modeling. Overall, this work provides a practical framework and a strong baseline for the digitization of historical manuscripts. While further validation on additional collections is necessary, extending the approach to other historical manuscript settings remains an important direction for future work. By releasing our code and models, we hope to support future research in historical document processing and contribute to ongoing efforts to preserve Nepal’s written heritage.

Limitations

While our approach demonstrates promising results, several limitations should be considered when applying this pipeline to the transcription of historical manuscripts. First, our study faces significant data constraints that limit the ability to assess generalizability. Although our data-centric approach partially mitigates this issue, our training corpus cannot fully capture the diversity of handwriting styles, document conditions, and formats found across historical archives. This limitation is particularly evident in the model’s degraded performance on text lines exceeding 120 characters, which we attribute to the scarcity of long-sequence examples in our training data. Second, the limited public availability of training data poses a broader challenge to the field. Most datasets used in this study cannot be released publicly due to copyright restrictions, thereby hindering both accessibility and collaborative development of handwritten text recognition systems for historical documents. Nevertheless, we believe that releasing our trained model represents a valuable contribution for scholars working with Old Nepali manuscripts. Future work could also examine optimization parameters, such as warmup steps, which have been shown to be effective in recent studies (Ströbel et al., 2023). Finally, our current pipeline relies on pre-segmented line inputs, which require external segmentation tools such as Kraken. This dependency introduces a potential source of error, particularly for documents with irregular layouts or severe physical damage. Future work should explore segmentation methods that adapt to the varied layouts commonly found in historical manuscripts.

Ethics Statement

We affirm that our research adheres to the [ACL Ethics Policy](#). This work does not involve human subjects or reveal any personally identifiable information. We declare that we have no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are acknowledged in the acknowledgments section. We have made every effort to document our methodology, experiments, and results accurately and are committed to sharing our code and other relevant resources to foster reproducibility and further research advances.

Acknowledgments

We would like to express our sincere gratitude to the colleagues in the research unit “Documents on the History of Religion and Law of Pre-modern Nepal” at the Heidelberg Academy of Sciences and Humanities for their assistance and for providing additional ground-truth data. We also thank Jan Kamlah of the University Library Mannheim for his constant support throughout this project. Finally, Esteban Garces Arias wishes to thank the Munich Center for Machine Learning (MCML) for their ongoing support.

References

- David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.
- Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. [Advancements and challenges in handwritten text recognition: A comprehensive survey](#). *Journal of Imaging*, 10(1).
- G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Edgard Chammas, Chafic Mokbel, and Laurence Likforman-Sulem. 2018. [Handwriting recognition of historical documents with few labeled data](#). In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 43–48.
- Simon Cubelic, Axel Michaels, and Astrid Zotter. 2018. [Towards a South Asian Diplomatics: Cosmopolitan Norms and Regional Idioms in the Use of Documents](#), pages 1–33. In *Studies in Historical Documents from Nepal and India, Documenta Nepalica: Book Series*. Heidelberg University Publishing.
- Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Gabriel Calazans Duarte de Moura, and Alejandro Héctor Toselli. 2024. [Data augmentation for offline handwritten text recognition: A systematic literature review](#). *SN Computer Science*, 5(2).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Yuanhao Ding, Esteban Garces Arias, Meimingwei Li, Julian Rodemann, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, and Chongsheng Zhang. 2025. [Guard: Glocal uncertainty-aware robust decoding for effective and efficient open-ended text generation](#). *Preprint*, arXiv:2508.20757.
- Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. 2026. [Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics](#). *Preprint*, arXiv:2604.11012.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). *Preprint*, arXiv:1805.04833.
- Markus Freitag and Yaser Al-Onaizan. 2017. [Beam search strategies for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.
- Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025. [Decoding decoded: Understanding hyperparameter effects in open-ended text generation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.
- Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. [Automatic transcription of handwritten old Occitan language](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.
- Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2024. [Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.
- Alex Graves and Jürgen Schmidhuber. 2008. [Offline handwriting recognition with multidimensional recurrent neural networks](#). In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.
- Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localization in natural images. In *European Conference on Computer Vision*, pages 231–246.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Internet Archive. 2025. Internet Archive. <https://archive.org>. Accessed: 2025-07-10.
- Benjamin Kiessling. 2019. [Kraken - A Universal Text Recognizer for the Humanities](#). In *Digital Humanities 2019*, Utrecht, Netherlands.
- Philipp Koch, Gilary Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin, and Matthias Assenmacher. 2023. [A tailored handwritten-text-recognition system for medieval Latin](#). In *Proceedings of the Ancient Language*

- Processing Workshop*, pages 103–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.
- Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.
- Ze Liu, Yutong Hu, Yixuan Lin, Zhicheng Lin, Zihang Gao, Ze Han, Xiang Chen, and et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.
- Nicole Merkel-Hilf. 2022. [Ground Truth data for printed Devanagari](#).
- Anthony Moi and Nicolas Patry. 2023. [HuggingFace’s Tokenizers](#).
- Swornim Nakarmi, Sarin Sthapit, Arya Shakya, Rajani Chulyadyo, and Bal Krishna Bal. 2024. [Nepal script text recognition using CRNN CTC architecture](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 244–251, Torino, Italia. ELRA and ICCL.
- Joseph Nockels, Paul Gooding, and Melissa Terras. 2024. [The implications of handwritten text recognition for accessing the past at scale](#). *Journal of Documentation*, 80(7):148–167.
- Nobuyuki Otsu. 1979. [A threshold selection method from gray-level histograms](#). *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.
- Mahes Raj Pant and Philip H. Pierce. 1989. *Administrative Documents of the Shah Dynasty Concerning Mustang and Its Periphery (1789–1844 A.D.)*, volume 10 of *Archiv für Zentralasiatische Geschichtsforschung*. VGH Wissenschaftsverlag, Bonn.
- John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalín, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps, and Franz Fischer. 2024. [Challenging error correction in recognised byzantine Greek](#). In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (MLAAL 2024)*, pages 1–12, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI.
- Yassin Hussein Rassul, Aram M. Ahmed, Polla Fattah, Bryar A. Hassan, Arwa W. Abdulkareem, Tarik A. Rashid, and Joan Lu. 2025. [Advancing offline handwritten text recognition: A systematic review of data augmentation and generation techniques](#). *Preprint*, arXiv:2507.06275.
- Theodore Jr. Riccardi. 1971. *A Nepali Version of the Vetālapañcaviṃśati*, volume 54 of *American Oriental Series*. American Oriental Society, New Haven.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). *Preprint*, arXiv:1508.07909.
- Baoguang Shi, Xiang Bai, and Cong Yao. 2017. [An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304.
- Raymond Smith. 2007. [An overview of the tesseract OCR engine](#). In *Proc. 9th Int. Conf. on Document Analysis and Recognition (ICDAR)*, pages 629–633.
- Mohamed Ali Souibgui, Alicia Fornés, Yousri Kessentini, and Beáta Megyesi. 2022. [Few shots are all you need: A progressive learning approach for low resource handwritten text recognition](#). *Pattern Recognition Letters*, 160:43–49.
- Phillip Benjamin Ströbel, Tobias Hodel, Walter Boente, and Martin Volk. 2023. [The Adaptability of a Transformer-Based OCR Model for Historical Documents](#), page 34–48. Springer Nature Switzerland.
- Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. [A contrastive framework for neural text generation](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561.
- Curtis Wigington, Seth Stewart, Brian Davis, Bill Barrett, Brian Price, and Scott Cohen. 2017. [Data augmentation for recognition of handwritten words and lines using a cnn-lstm network](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, page 639–645. IEEE.
- Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price, and Scott Cohen. 2018. [Start, follow, read: End-to-end full-page handwriting recognition](#). In *Computer Vision – ECCV 2018*, pages 372–388, Cham. Springer International Publishing.
- Christof Zotter. 2018. [Ascetics in Administrative Affairs: Documents on the Central Overseers of Jogīs and Saṃnyāsīs in Nepal](#), page 445–491. In *Studies in Historical Documents from Nepal and India, Documenta Nepalica: Book Series*. Heidelberg University Publishing.
- Rājendra Vikrama Śāha. 1832. *A lālamohara granting Munsī Lakṣmīdāsa a piece of land at Madu Tola as sunābirtā*. Heidelberg: HAdW (Documenta Nepalica). VS 1889. Edited and translated by Manik Barcharya. Published in 2017.

Appendix

A Parameters for image binarization

Method	Parameter
Otsu’s Binarization	Threshold = determined automatically by cv2.THRESH_OTSU

Table 5: Parameters for binarizing grayscale images (Otsu, 1979; Bradski, 2000).

B Effect of Binarization on Model Performance

Method	CER	Used in training
Binarized	0.098	No
Non-Binarized	0.089	Yes

Table 6: CER with and without binarization. Results are computed using three-stage training on the original Old Nepali dataset. Otsu’s global thresholding is used for binarization (Otsu, 1979). The best results are highlighted in **bold**.

C Model Architecture Comparison

Encoder	Decoder	Tokenizer	Training stage	Train time (hrs)	Eval runtime (s)	CER	CER(w)	ACC
trocr-base-handwritten	BERT	byteBPE	Pretraining w/ synthetic data	2.45	344.77	0.005	0.006	90.6%
			Transfer Learning w/ printed Nagari scripts	0.40	60.31	0.017	0.019	68.2%
			Final model training w/ Old Nepali scripts	0.69	61.66	0.082	0.082	24.8%
		charBPE	Pretraining w/ synthetic data	2.35	268.18	0.007	0.007	90.5%
			Transfer Learning w/ printed Nagari scripts	0.37	47.72	0.017	0.018	70.7%
			Final model training w/ Old Nepali scripts	0.56	39.88	0.087	0.086	25.5%
	GPT-2	byteBPE	Pretraining w/ synthetic data	2.13	282.74	0.007	0.007	90.9%
			Transfer Learning w/ printed Nagari scripts	0.39	51.12	0.018	0.019	66.1%
			Final model training w/ Old Nepali scripts	0.65	54.26	0.084	0.084	26.1%
		charBPE	Pretraining w/ synthetic data	2.34	217.95	0.007	0.009	90.7%
			Transfer Learning w/ printed Nagari scripts	0.36	42.41	0.017	0.017	72.7%
			Final model training w/ Old Nepali scripts	0.54	34.37	0.084	0.082	28.7%

Table 7: Results for all model combinations after the 3-stage training. Model selection is performed based on the CER metric. The best results are highlighted in **bold**. All experiments were performed with an NVIDIA GeForce RTX 4090 GPU. We use the dataset with normalized transcriptions to reduce variance across runs and ensure a stable comparison.

D Performance Metrics for Swin Encoder

Encoder variant	CER	ACC
trocr-base-handwritten	0.056	29.35%
swin-base-patch4-window7-224-in22k	0.174	21.93%

Table 8: CER and accuracy comparison between TrOCR and Swin encoders on the Old Nepali dataset. TrOCR outperforms Swin in both metrics for the same data and training configurations. The best results are highlighted in **bold**.

E Decoder Benchmarking

Decoder	CER	Total model parameters	Evaluation (samples/sec)
TrOCR decoder	0.059	334M	3.50
Script-aware decoder (ours)	0.056	202M	4.96

Table 9: Comparison of the standard TrOCR decoder (trocr-base-handwritten) with our script-aware decoder (with a BERT architecture, byteBPE tokenizer, vocab size = 500), under identical configurations using the normalized transcriptions and $8\times$ augmented dataset for fair comparison. The best results are highlighted in **bold**.

F Data Statistics

Statistic	Value
<i>Image-Level Statistics</i>	
Number of manuscripts	155
Approximate time period	18th and 19th centuries
Mean characters per script	1,198
Mean lines per script	20
<i>Line-Level Statistics</i>	
Total cropped lines	3,100
Mean characters per line	60
Shortest line length	1 character
Longest line length	162 characters
Standard deviation of line length	27.65
Average image size (width \times height)	1593 \times 133 pixels

Table 10: Table showing detailed statistics of the dataset at full image (manuscript) and line level.

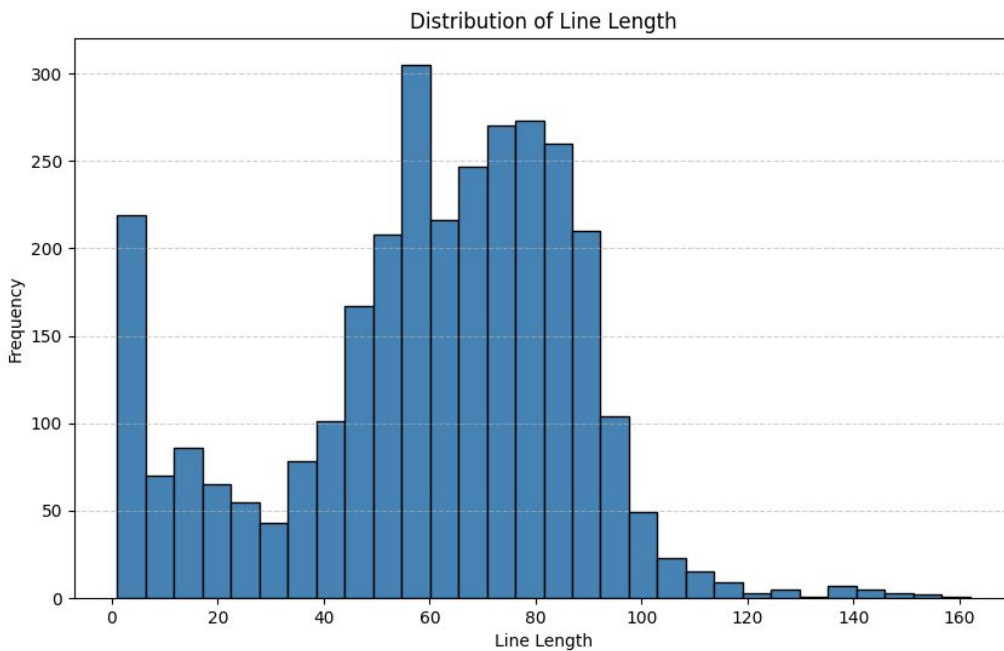


Figure 7: Histogram showing the distribution of the line lengths across all text samples in the dataset.

G Character Inventory

#	Character	Frequency	Relative frequency (%)	#	Character	Frequency	Relative frequency (%)
1	।	19516	11.5030	2	॰	13258	7.8145
3	र	11403	6.7211	4	न	7352	4.3334
5	ि	7352	4.3334	6	क	6798	4.0068
7	म	5749	3.3885	8	व	5575	3.2860
9	त	5447	3.2105	10	स	5025	2.9618
11	॰	4768	2.8103	12	य	4767	2.8097
13	ल	4726	2.7856	14	ह	4281	2.5233
15	ो	4158	2.4508	16	॰	4153	2.4478
17	॰	3997	2.3559	18	प	3849	2.2687
19	द	3833	2.2592	20	ी	3527	2.0789
21	ग	3294	1.9415	22	ज	3192	1.8814
23	भ	2194	1.2932	24	ष	1902	1.1211
25	॰	1783	1.0509	26	श	1536	0.9053
27	॰	1461	0.8611	28	छ	1166	0.6873
29	च	1092	0.6436	30	॰	966	0.5694
31	आ	933	0.5499	32	झ	905	0.5334
33	उ	886	0.5222	34	ध	868	0.5116
35	अ	867	0.5110	36	झ	799	0.4709
37	ट	796	0.4692	38	॰	699	0.4120
39	फ	691	0.4073	40	थ	678	0.3996
41	ड	673	0.3967	42	ण	643	0.3790
43	॰	569	0.3354	44	१	562	0.3313
45	ठ	491	0.2894	46	॥	426	0.2511
47	।	376	0.2216	48	ी	310	0.1827
49	ञ	306	0.1804	50	२	275	0.1621
51	घ	259	0.1527	52	०	198	0.1167
53	३	192	0.1132	54	८	187	0.1102
55	ए	183	0.1079	56	॰	177	0.1043
57	ढ	176	0.1037	58	५	174	0.1026
59	ब	147	0.0866	60	४	136	0.0802
61	झ	124	0.0731	62	:	123	0.0725
63	६	121	0.0713	64	ख	116	0.0684
65	९	112	0.0660	66	७	101	0.0595
67	:	82	0.0483	68	औ	69	0.0407
69	\	43	0.0253	70	ऊ	37	0.0218
71	ड	35	0.0206	72	ऐ	24	0.0141
73	/	18	0.0106	74	ऋ	13	0.0077
75	॰	13	0.0077	76	ओ	7	0.0041
77	-	5	0.0029	78	५	4	0.0024
79	(1	0.0006	80)	1	0.0006

Table 11: Character frequency and relative frequency observed in the ground truth labels.

H Training Pipeline

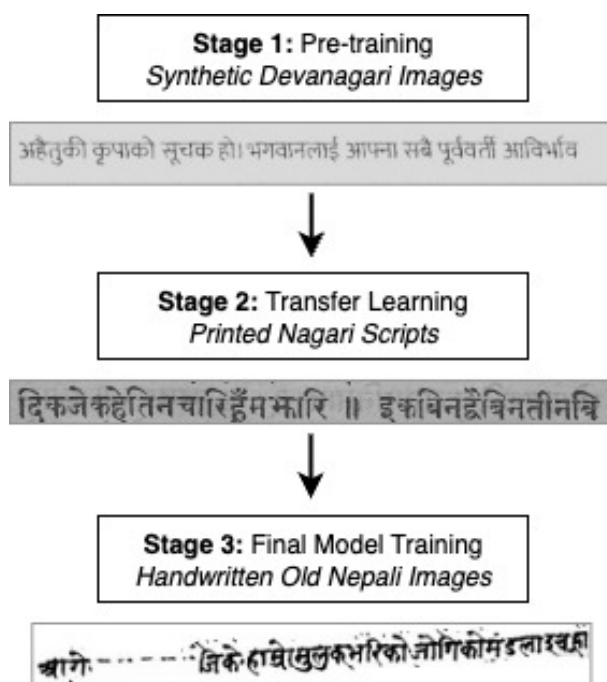


Figure 8: Visualization of the three-stage training pipeline with an example image from each dataset.

I Transcription Normalization Statistics

Normalization step	Count	Lines affected	% of lines
Bullet symbols	4,965	1,780	57.41%
Total number of extra space fixed	182	129	4.1%
Removing upper dashes (combining macrons)	150	87	2.8%
Pipe to danda conversion	3	3	<0.1%
Convert ASCII digits to Devanagari digits	4	3	<0.01%
Standardized chandrabindu	3	3	<0.1%

Table 12: Summary of character-level normalization steps applied during label cleaning. Each operation counts the number of characters that were modified or removed.

J Augmentation Methods

Augmentation	Description
<i>Shape and Angle Distortions</i>	
rotation	Rotate the image by a small angle ($\pm 3^\circ$)
shift	Shifts the whole image randomly by 5-20 pixels in the x and y directions
perspective	Tilts the image to simulate a camera picture taken from an angle
shear	Skews the image diagonally to mimic slanted text
hstretch	Stretches the image horizontally by a factor of 1.2
vstretch	Stretches the image vertically by a factor of 1.2
<i>Noise and Blur Artifacts</i>	
blur	Applies Gaussian blur with kernel size $5 \times$ across the whole image
motion_blur	Blurs in one direction to mimic movement during scanning
jpeg	Simulates quality loss from JPEG compression
jitter	Adds random pixel-level noise
gaussian_noise	Applies Gaussian noise to the full image, which adds random bright or dark speckles
elasticblur	Combines elastic distortion with Gaussian blur
saltpepper	Adds black and white specks like dust or scan noise
<i>Letter-Level Damage Simulation</i>	
blurredpatches	Picks random spots on the image and blurs to simulate smudges
sine	Applies a wave horizontal distortion using sine curves
horizontal	Curves the text horizontally, like bending the page gently
elastic	Bends and wraps the image using smooth distortion, like crumpled paper
med_blur	Applies a light median filtering with kernel size 3×3 , to smoothen very small irregularities
morph	Applies erosion or dilation to simulate ink spread or fade
sharpen	Enhances edge contrast using a 3×3 convolutional kernel to simulate bold ink or high-contrast scanning

Table 13: List of image-level augmentations used for generating synthetic data of Old Nepali image samples. All the augmentations are applied to grayscale, non-binarized images to preserve original stroke quality and noise characteristics.

K Synthetic Dataset Preparation Fonts and Noise

Fonts used for synthetic data generation

Fonts	Source
Akasha Regular	Link
Anek Devanagari	Google Fonts
Chandas	Link
IBM Plex Sans Devanagari	Google Fonts
Kalam	Google Fonts
Lohit Devanagari	Link
Mukta	Google Fonts
Noto Sans Devanagari	Google Fonts
Palanquin Dark	Google Fonts
Siddhanta	Link
Tiro Devanagari Hindi	Google Fonts

Table 14: Fonts used to generate the synthetic dataset (first stage). These fonts were selected to reflect a range of Devanagari styles.

Augmentations used for adding noise to synthetic data generation

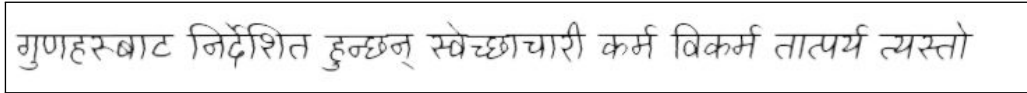
Augmentation	Parameters
Piecewise Affine	$p = 0.6$, scale = (0.005, 0.015)
Elastic Transformation	$p = 0.5$, alpha = (1.5, 3.0), sigma = (0.6, 1.0)
Motion Blur	$p = 0.4$, kernel size $k = [3, 5]$
Dropout	$p = 0.4$, dropout rate = (0.01, 0.03)
Gaussian Blur	$p = 0.4$, sigma = (0.5, 0.9)
Linear Contrast	$p = 0.3$, factor = (0.7, 1.4)
Brightness Scaling	$p = 0.3$, multiply factor = (0.85, 1.15)
Laplace Noise	$p = 0.3$, scale = 0.01
Vertical Translation	$p = 0.5$, shift = $\pm 1.5\%$
Convolution Blur	3×3 smoothing kernel

Table 15: Augmentation operations used during synthetic dataset generation. p refers to the probability of applying each transformation. Each operation is applied probabilistically using modules from the `imgaug` library.

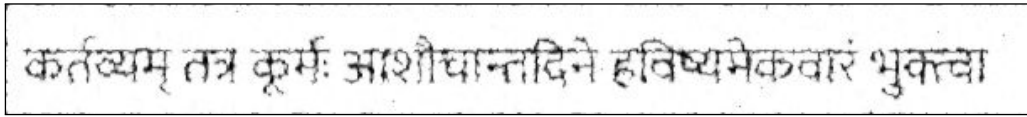
Text sources for generating Old Nepali synthetic dataset

Name	Source
Aatma Gyan	https://archive.org/details/aatma-gyan/page/n21/mode/2up
Antya Karma Paddhati in Nepali	https://archive.org/details/antya-karma-paddhati-pdf
Bhagwat Gita in Nepali	https://archive.org/details/20220411_20220411_0657-bhagavad-gita-in-nepali
Bhanubhakta Ramayan	https://archive.org/details/nepali-bhanubhakta-ramayan
Chanakya Niti in Nepali	https://archive.org/details/chanakya-niti-in-nepali_hinduism
Devkala Vamshavali (Yogi Naraharinath)	https://archive.org/details/DevkalaVamshavaliYogiNaraharinath
Ekadeshma	https://archive.org/details/ekadeshma_202204
Gorkha Vamshavali	https://archive.org/details/gorkha-vamshavali
Grhapravesa	https://archive.org/details/grhapravesa-aitihāsika-upanyasa
Laxmi Prasad Devkota 2025 BS Shakuntala Mahakavya	https://archive.org/details/laxmi-prasad-devkota-2025-bs-shakuntala-mahakavya
Madhyacandrik	https://archive.org/details/madhyacandrik00sharuoft/page/n5/mode/2up
Mahabharat in Nepali	https://archive.org/details/mahabharat-sampoorna-18-parva-nepali
Maharani Rajyalakshmi	https://archive.org/details/kotparva-ki-maharani-rajyalakshmi
Muna Madan	https://archive.org/details/MunaMadan
Naridharma tatha Purushartha	https://archive.org/details/20220206_20220206_1126
Nepal Ritupau 1956	https://archive.org/details/nepal-ritupau-1956_202204
Nepal ko itihās	https://archive.org/details/428977-nepal-ko-itihās-1950
Prameshwar Adalat	https://archive.org/details/Prameshwar-ko-adalat-ma-veda-ra-bible
Rig Ved in Nepali Language	https://archive.org/details/rig-ved-in-nepali_202112
Shreemad Bhagwat Mahapurān in nepali	https://archive.org/details/shreemad-bhagwat-mahapurān
Teachings of Queen Kunti	https://archive.org/details/20220412_20220412_1153-teachings-of-queen-kunti

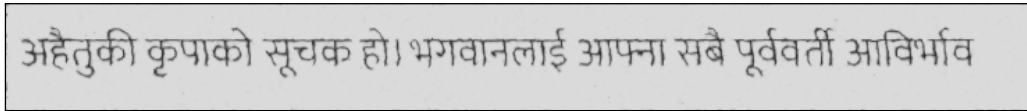
Table 16: Text sources for generating Old Nepali synthetic corpus and synthetic images.



(a) Sample 1



(b) Sample 2



(c) Sample 3

Figure 9: Three samples of synthetic Devanagari line images generated from the text corpus, which is used in the first stage of training.

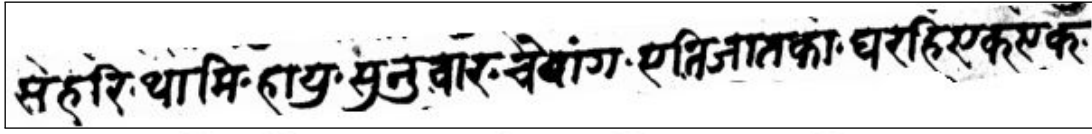
L Decoding Methods and Results

Method	Hyperparameters	Weighted CER	Mean CER
Beam Search	$width = 1$	0.0483	0.0577
Beam Search	$width = 5$	0.0483	0.0577
Beam Search	$width = 10$	0.0483	0.0577
Beam Search	$width = 20$	0.0483	0.0577
Contrastive Search	$k = 5, \alpha = 0.2$	0.0488	0.0581
Contrastive Search	$k = 5, \alpha = 0.6$	0.0488	0.0581
Contrastive Search	$k = 5, \alpha = 0.8$	0.0488	0.0581
Contrastive Search	$k = 10, \alpha = 0.2$	0.0488	0.0581
Contrastive Search	$k = 10, \alpha = 0.6$	0.0488	0.0581
Contrastive Search	$k = 10, \alpha = 0.8$	0.0488	0.0581
Temperature Sampling	$\tau = 0.2$	0.0488	0.0581
Temperature Sampling	$\tau = 0.4$	0.0488	0.0581
Temperature Sampling	$\tau = 0.6$	0.0488	0.0581
Temperature Sampling	$\tau = 0.8$	0.0488	0.0581
Temperature Sampling	$\tau = 0.9$	0.0485	0.0579
Temperature Sampling	$\tau = 1.0$	0.0488	0.0580
Top-k Sampling	$k = 3$	0.0488	0.0571
Top-k Sampling	$k = 5$	0.0487	0.0580
Top-k Sampling	$k = 10$	0.0489	0.0582
Top-k Sampling	$k = 20$	0.0486	0.0579
Top-k Sampling	$k = 50$	0.0490	0.0582
Top-p Sampling	$p = 0.5$	0.0488	0.0580
Top-p Sampling	$p = 0.6$	0.0488	0.0580
Top-p Sampling	$p = 0.7$	0.0488	0.0580
Top-p Sampling	$p = 0.8$	0.0488	0.0580
Top-p Sampling	$p = 0.9$	0.0488	0.0581
Top-p Sampling	$p = 0.95$	0.0488	0.0580

Table 17: Comparison of decoding methods and hyperparameter configurations. All results are reported using the standard model setup with a `trocr-large-handwritten` encoder, `byteBPE` tokenizer, and `script-aware BERT` decoder. Decoding strategies and hyperparameters are based on sensitivity analysis by [Garces Arias et al. \(2025\)](#).

M Benchmarking Analysis

Input sample segmented from Figure 17.



सहरि.थामि.हायु.सुनुवार.चेपांग.एतिजातका.घरहिएकएक

(a) Input sample of manuscript line (top) with ground truth (bottom).



(b) Prediction from fine-tuned TrOCR baseline model.



(c) Prediction from Google Cloud Vision API.

Figure 10: Benchmarking examples showing input and outputs from two OCR baselines. The letters highlighted in red are incorrect predictions.

N Evaluating the effect of each training stage

Training stage	Pretraining	CER	ACC
Stage 1	-	0.71	0.0%
Stage 2	Stage 1	0.51	2.58%
Stage 3	Stage 1 + Stage 2	0.056	29.58%

Table 18: Effect of each training stage on model performance (CER and accuracy), showing how performance improves from Stage 1 to Stage 3 when evaluated on the final (Stage 3) test set.

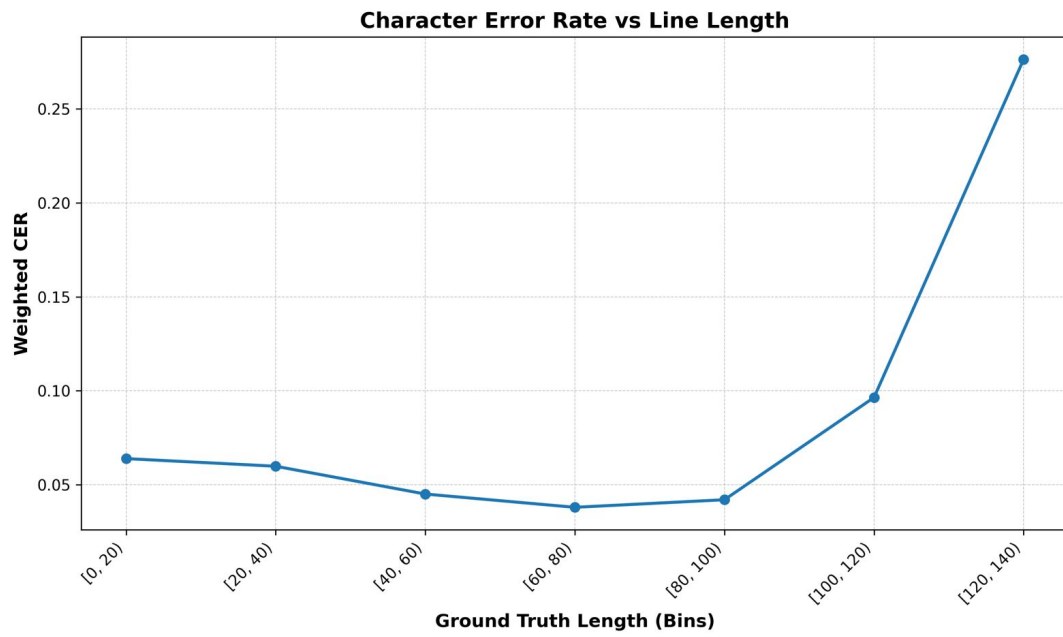


Figure 13: Evaluation of CER against the total number of characters in a line. Each bin groups lines by character count, and the CER is averaged using line lengths as weights.

Most common errors

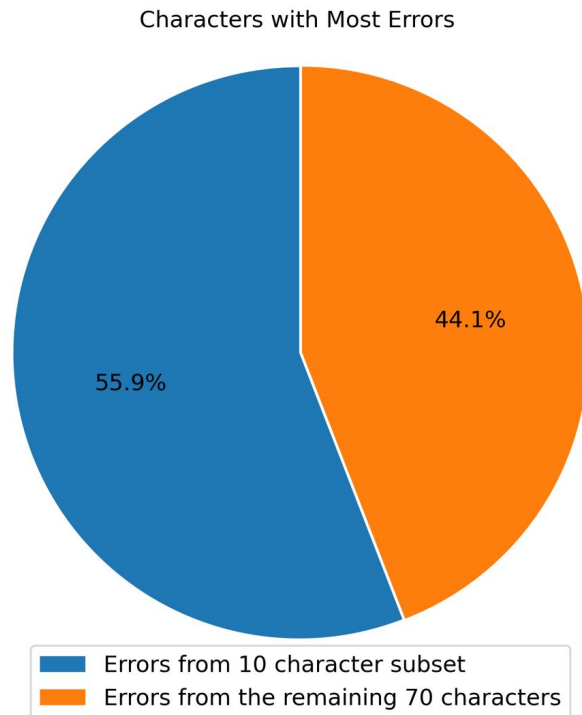


Figure 14: Error share between the top-10 characters with most errors and the remaining 70 characters. The error share is calculated as the ratio of total errors by characters to the total errors in the test set.

Character	Name	Error count	Error share
˘	VIRAMA	111	12.92%
	SPACE	98	11.41%
र	CONSONANT RA	56	6.52%
ॄ	VOWEL U	35	4.07%
·	SIRBINDU	34	3.96%
·	NUKTA	33	3.84%
.	FULL STOP	33	3.84%
।	VOWEL AA	33	3.84%
न	CONSONANT NA	25	2.79%
प	CONSONANT PA	23	2.68%

Table 19: Overview of most common errors with the count of errors in the test set, along with its error share. The total error count is 859.

P Token Uncertainty Analysis

We analyze whether the model’s own uncertainty can be used to flag incorrect predictions. To do this, we compute the relative probability of each predicted token as:

$$\text{Relative Probability} = \frac{\text{prob}_2}{\text{prob}_1}$$

Here, prob_1 and prob_2 refer to the probabilities of the top-1 and top-2 predicted tokens, respectively. These probabilities are extracted from the decoder’s output logits, to which we then apply a softmax. We obtain the probability distribution over the vocabulary and then record the probabilities. A high relative probability indicates that the model is uncertain, as it suggests possible confusion between the top two predictions. We therefore assess whether this metric can detect incorrect tokens. We evaluate the performance of this score in distinguishing correct from incorrect tokens using precision, recall, and F1 score. The best F1 score is achieved at a threshold of 0.034 (Table 20). This threshold is then used to flag low-confidence predictions in further analysis.

Metric	Score
F1-score	0.306
Precision	0.363
Recall	0.264

Table 20: Metrics for detecting incorrect tokens using the model’s relative probability between top predictions.

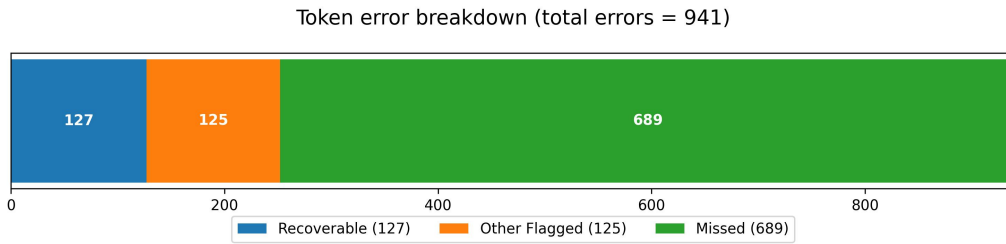


Figure 15: Breakdown of 941 token-level errors based on relative probability. Out of these, 236 tokens were flagged by the model as low-confidence, and 127 of them were recoverable, i.e., the correct token appeared in the top-3 predictions, making them useful for post-correction or human review.

Q Performance Gains

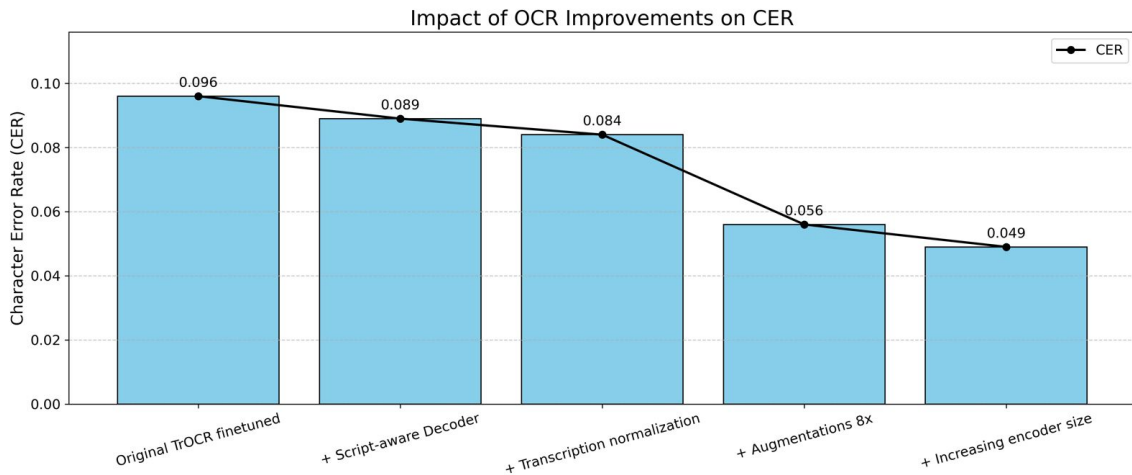


Figure 16: Stepwise reduction in CER through successive improvements to the HTR pipeline.

R Manuscript Samples

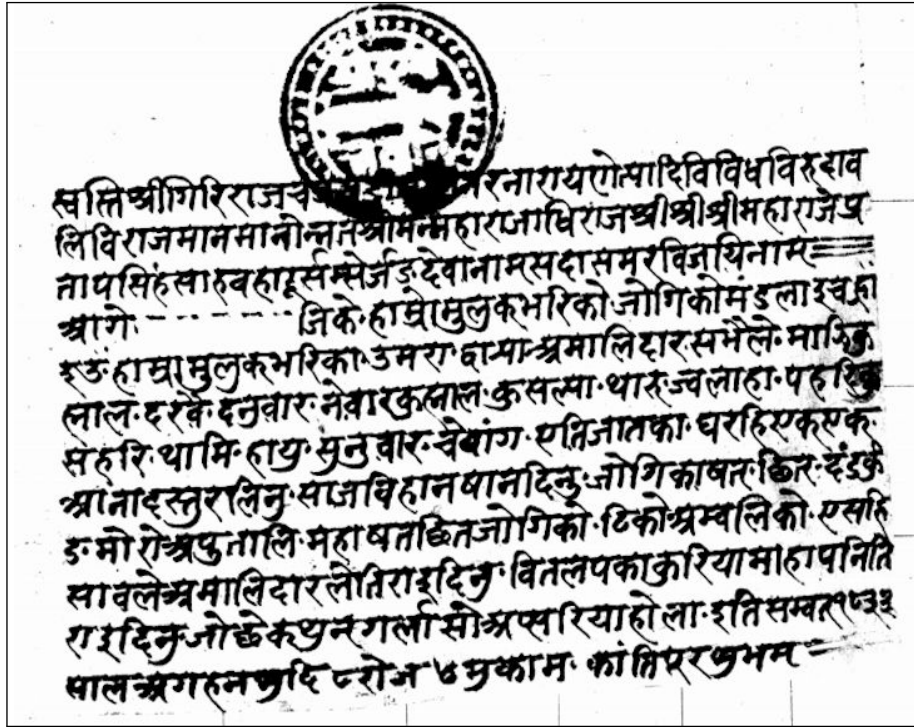


Figure 17: Manuscript image of NGMPP DNA 14/50 (© National Archives Nepal) excerpted and cropped from (Zotter, 2018). Lines from this paragraph were segmented and used in our OCR dataset, with 80% allocated to training, 10% to validation, and 10% to testing. This figure illustrates a representative sample of our dataset.

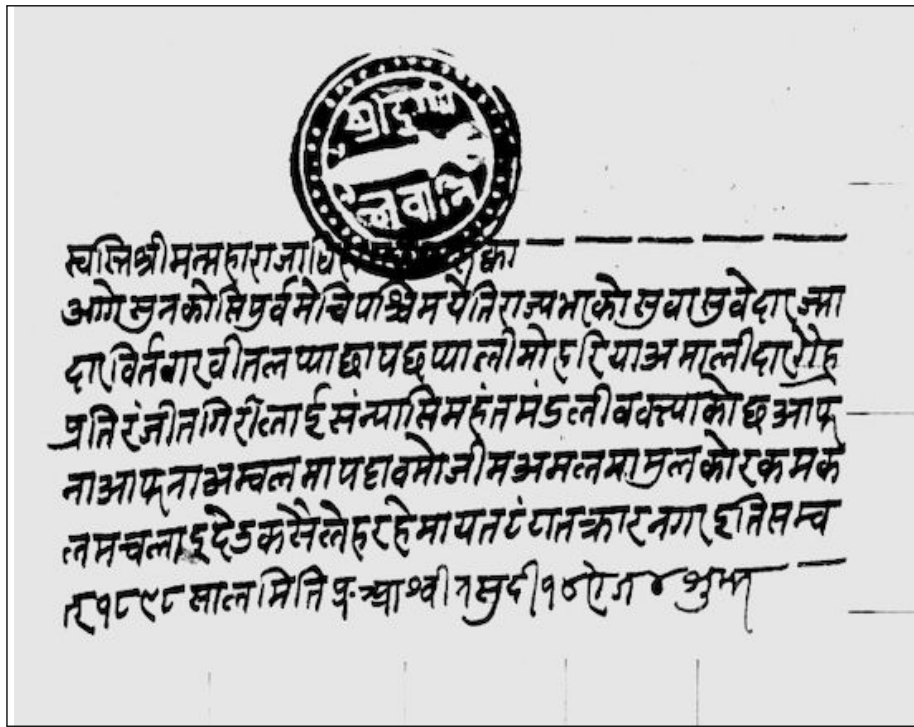
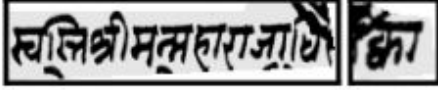


Figure 18: Another sample of a manuscript image (NGMPP DNA 13/59 © National Archives Nepal) cropped from (Zotter, 2018), also included in our training dataset.

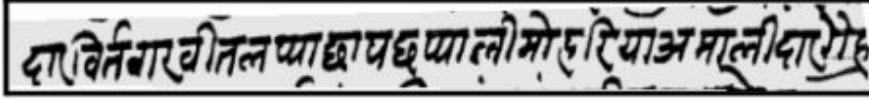
Line 1:



Line 2:



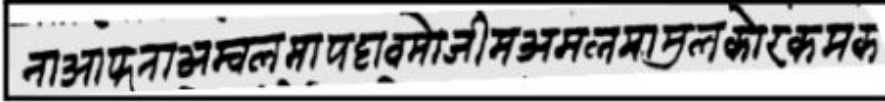
Line 3:



Line 4:



Line 5:



Line 6:



Line 7:

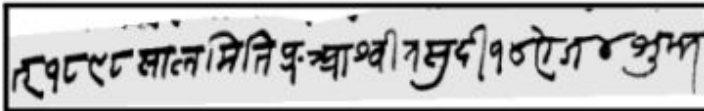


Figure 19: Line-level segmentation of Figure 18, performed using Kraken's polygon based method. The text region is segmented into 7 lines, which are used as training samples.

Input sample segmented from Figure 17

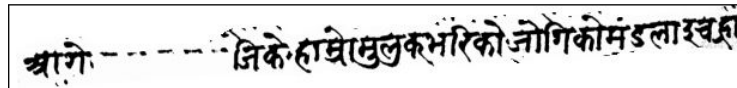


Figure 20: This image displays the upper dashed line ambiguity, which is normalized to reduce stylistic noise.

S Interactive App

Interactive segmentation using Kraken's polygon method

app

- segmentation
- prediction
- analysis

1. Input & Segmentation

Upload image (JPEG/PNG)

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

Browse files

Run Segmentation

Choose options

To find the line numbers, hover over the lines in the plot below.

Line-by-line HTR prediction on the segmented lines

Segmentation Overlay

OCR Predictions

Which line(s) to OCR?

All

Run OCR & Show

- स्वस्तिश्रीगिरिराजचक्रचूडामणिनरनारायणेत्यादिविधिविरुदावलीविरा
- जमानमानोत्रतश्रीमन्महाराजाधिराजश्रीश्रीमहाराजराजेन्द्रविक्रमसाव
- वहादूरसम्प्रेरजडधेवानांसदासमरविजयीनाम
- आगोसहरकाठमाडौंगडटोलमडिठिकोडेविदेघरकोमुत्सीलशिषदासके.सो
- हिटोलसोहीजोनिघरकोधनरतेजनाराजकावावुधंचालेगुठिमाराशिययाको
- सोहीघरमतिकोवारिपाताल.तस्कोसांध.जेरापुरानावारिकापषलदेषिपा
- श्रिमःदमाईकोवारिदेधिउन्तर.तैलेजनारानसंगकिन्याकोवारिदेधिपुर्वदि
- लक्षिनरसिसंगकिन्याकोघरकावल्यानिपातालदेषिदछिनःयेतिसाधमि
- अवारिपातालषागुहात18अंगुल3काआफ्नाविर्ताकोषादोलषेतरोपनीउ
- ई2मथ्येरोपनी1दि.आफ्ना2मनोमानसंगःगुठिपातालकाषेतसद्दादिःपाता
- लसिराष्याकोसोहीवमोजीसर्महामिले.मोहोइगरिवकस्यो.तस्कोवकसाडलीहै