NE-OCR

Unified Optical Character Recognition for the Languages of Northeast India

MWire Labs Technical Report | March 2026

MWire Labs Research
Shillong, Meghalaya, India
connect@mwirelabs.com

Model: huggingface.co/MWirelabs/ne-ocr
License: CC-BY-4.0

Abstract

We present NE-OCR, a unified optical character recognition model for 10 Northeast Indian languages - represented across 12 language-script pairs spanning 4 scripts - along with Hindi and English as anchor languages. NE-OCR is built on a Vision Transformer backbone (ViTSTR-Base, 86M parameters), trained on approximately 1.34 million text-image pairs constructed from native language corpora. On a held-out benchmark of 24,000 test samples (2,000 per language-script pair), NE-OCR achieves a mean Character Accuracy (ChA) of 94.99%, reaching a peak of 98.85% on Khasi, while maintaining an inference latency of 17.2ms per image on an A40 GPU - the fastest among all evaluated systems. We benchmark against four baseline systems: EasyOCR, Tesseract 5, TrOCR-large-printed, and Chandra. NE-OCR outperforms all baselines across 9 Northeast Indian language-script pairs, with competitive performance on the English and Hindi anchor languages. We additionally present a qualitative analysis of DeepSeek-OCR 2 and Chandra as representatives of the vision-language model (VLM) paradigm, demonstrating that VLMs fail on unseen scripts by hallucinating document structure rather than producing recognition errors. Model weights are publicly available under CC-BY-4.0.

Contents

1. Introduction

Northeast India is home to over 220 documented languages from four major language families - Tibeto-Burman, Indo-Aryan, Austroasiatic, and Tai-Kadai - spoken across eight states by approximately 45 million people. Despite this linguistic richness, the region remains severely underserved by existing language technology infrastructure. Standard OCR systems, designed primarily for high-resource languages, fail to generalize to the typographic and script diversity of the Northeast.

The practical consequences are significant. Government documents, educational materials, historical archives, and digital newspapers in languages such as Khasi, Mizo, Kokborok, Bodo, and Meitei Mayek cannot be reliably digitized using off-the-shelf tools. In our evaluation, EasyOCR achieves 2.50% Character Accuracy on Meitei Mayek text - effectively producing random output. Tesseract 5 achieves 2.24% on the same script, with a Word Error Rate of 106.46%, meaning it inserts more words than are present in the ground truth. TrOCR-large-printed, a strong model for English printed text, achieves near-zero accuracy on all non-Latin scripts.

MWire Labs has been developing foundational language technology for Northeast India since 2024, including large language models, speech systems, and document processing tools. NE-OCR is part of this broader effort to build a comprehensive language technology stack for the region. In this report, we describe the development of a unified OCR model covering 10 Northeast Indian languages across 4 scripts, along with Hindi and English as anchor languages, trained on large-scale data constructed specifically for this purpose.

The key contributions of this work are:

1. A unified OCR model covering 10 Northeast Indian languages (12 language-script pairs, 4 scripts), released publicly under CC-BY-4.0.
2. A training corpus of approximately 1.34 million text-image pairs constructed from native language corpora using language-appropriate fonts and augmentation pipelines.
3. A standardized benchmark of 24,000 test samples (2,000 per language-script pair) with Character Accuracy, Character Error Rate, and Word Error Rate.
4. A comprehensive comparison against EasyOCR, Tesseract 5, TrOCR-large-printed, and Chandra, showing NE-OCR achieves best performance on 9 language-script pairs.
5. A qualitative analysis of VLM-based OCR failure modes on unseen scripts, using DeepSeek-OCR 2 and Chandra as case studies.

2. Related Work

2.1. Scene Text Recognition

Optical character recognition has advanced considerably with deep learning. The CRNN framework (Shi et al., 2016) established the encoder-decoder paradigm combining convolutional feature extraction with BiLSTM sequence modeling and CTC decoding. Subsequent work explored attention-based decoders,

spatial transformation networks, and transformer architectures. Atienza (2021) proposed ViTSTR, applying Vision Transformers (Dosovitskiy et al., 2020) directly to the text recognition task as a single-stage model, demonstrating competitive accuracy with significantly fewer parameters. NE-OCR adapts ViTSTR-Base as implemented in the DocTR library.

TrOCR (Li et al., 2021) applies an encoder-decoder transformer - a ViT encoder with a language model decoder - to printed and handwritten text recognition, achieving state-of-the-art results on English benchmarks. However, as our evaluation demonstrates, TrOCR-large-printed fails completely on non-Latin scripts, with near-zero ChA on Assamese, Hindi, Bodo, and Meitei Mayek.

2.2. Multilingual and Low-Resource OCR

General-purpose multilingual OCR systems such as EasyOCR, Tesseract, PaddleOCR, and Surya have expanded language coverage considerably. However, coverage of low-resource languages - particularly those with non-Latin scripts and limited digital corpora - remains inconsistent. Tesseract requires language-specific trained data files that are unavailable for most Northeast Indian languages. EasyOCR's recognition module relies on CRNN-based models trained on publicly available datasets that provide limited coverage of Northeast scripts.

Recent work on low-resource OCR has explored synthetic data generation as a primary training strategy, particularly for scripts with limited annotated document collections. The Mozhi Indic OCR dataset (Mathew et al., 2025) provides real scanned training data for Assamese, Hindi, and Manipuri. NE-OCR extends this approach to languages not covered by existing resources, and introduces a unified multi-script training regime across all language families of the region.

2.3. Vision-Language Models for OCR

Large vision-language models (VLMs) such as GPT-4V, Gemini, Qwen-VL, and DeepSeek-VL2 have demonstrated impressive general OCR capabilities on English and high-resource language documents. However, their performance on unseen scripts and low-resource languages has not been systematically studied. In this report, we provide a preliminary evaluation of DeepSeek-OCR 2 and Chandra on the NE-OCR benchmark, observing that both models fail on Meitei Mayek and Bengali-script NE languages, producing hallucinated document structure rather than recognition errors. This finding motivates the continued development of script-specific trained models for the region.

2.4. Language Technology for Northeast India

Academic and industrial work on NLP for Northeast India has grown but remains concentrated on a small subset of languages, primarily Assamese, Meitei, and Mizo. MWire Labs has been developing foundational language models for Northeast Indian languages. NE-OCR complements this work by addressing the document digitization layer of the language technology stack.

3. Benchmark

3.1. Languages, Scripts, and Dataset Framing

The NE-OCR benchmark covers 10 Northeast Indian languages represented across 12 language-script pairs, spanning 4 scripts. Meitei is represented twice: once in its native Meitei Mayek script and once in Bengali script, reflecting the dual-script reality of written Meitei in practice. Hindi and English are included as anchor languages to calibrate model performance against widely benchmarked baselines. Table 1 summarizes the training data composition and test set size per language-script pair.

Language	Script	Train Samples	Test Samples
Mizo	*Latin*	~120k	2,000
Garo	*Latin*	~110k	2,000
Nyishi	*Latin*	~100k	2,000
Khasi	*Latin*	~136k	2,000
Kokborok	*Latin*	~109k	2,000
Nagamese	*Latin*	~91k	2,000
Bodo	*Devanagari*	~105k	2,000
Meitei (Meitei Mayek)	*Meitei Mayek*	~86k	2,000
Assamese	*Bengali*	~100k	2,000
Meitei (Bengali script)	*Bengali*	~100k	2,000
Hindi (anchor)	*Devanagari*	~100k	2,000
English (anchor)	*Latin*	~100k	2,000
NE Place Names	*Latin*	~87k	-

Table 1: NE-OCR training and test set composition by language-script pair. 'Anchor' languages (Hindi, English) are included for calibration.

3.2. Test Set Construction

The test set consists of 24,000 text-image samples, with 2,000 samples per language-script pair. Test samples are drawn from held-out text segments comprising real printed/scanned images (≈38% of the overall data mix) as well as synthetically rendered text not seen during training.

3.3. Evaluation Metrics

We report three metrics across all language-script pairs and models:

- **Character Accuracy (ChA%):** Percentage of characters correctly predicted. Higher is better.
- **Character Error Rate (CER%):** Levenshtein edit distance at character level, normalized by ground truth character count. Lower is better. Values above 100% indicate the model produces more characters than the ground truth.
- **Word Error Rate (WER%):** Levenshtein edit distance at word level, normalized by ground truth word count. Lower is better. Values above 100% indicate insertion of more words than exist in the reference.

We additionally report inference latency in milliseconds per image, measured on an NVIDIA A40 48GB GPU.

4. Model Architecture

4.1. Overview

NE-OCR is built on the ViTSTR-Base architecture (Atienza, 2021) as implemented within the DocTR OCR framework. ViTSTR applies a Vision Transformer encoder directly to fixed-size image patches, producing patch-level representations that are decoded into character sequences via CTC loss. The single-stage CTC formulation eliminates the attention alignment bottleneck of encoder-decoder models and provides strong parallelism at inference time - a key advantage for deployment across the 12 language-script pairs covered by NE-OCR.

The architecture was selected after evaluating alternatives during development. Florence-2, a large vision-language model, was explored for per-language fine-tuning but produced poor results on Latin-script Northeast languages, achieving 59–77% ChA on Mizo and 63–64% on Nyishi, due to its language model decoder being biased toward English character distributions. The DocTR ViTSTR-Base architecture, trained jointly across all languages and scripts, consistently outperformed per-language Florence-2 fine-tunes and was adopted as the final architecture.

4.2. Architecture Details

The model processes RGB images resized to 32 × 128 pixels (height × width), normalized to the [0, 1] range and formatted as CHW tensors. No additional preprocessing steps are applied. Images are divided into non-overlapping patches of size 4 × 8 pixels and projected to a fixed embedding dimension. Learnable position encodings are added to patch embeddings before being passed through 12 transformer encoder layers with multi-head self-attention and feed-forward sublayers. The final encoder representations are projected to a vocabulary of 1,056 characters via a linear head, and CTC decoding with beam search produces the output character sequence.

The vocabulary encompasses characters from all four target scripts: Latin (including extended characters specific to Khasi, Kokborok, and Mizo orthographies), Devanagari (Hindi and Bodo), Bengali (Assamese and Meitei in Bengali script), and Meitei Mayek. It was constructed by taking the union of the DocTR multilingual character set (726 characters) and all characters present in the NE language training corpora, yielding 1,055 characters plus the CTC blank token at index 0.

[Figure 1: NE-OCR Model Architecture]

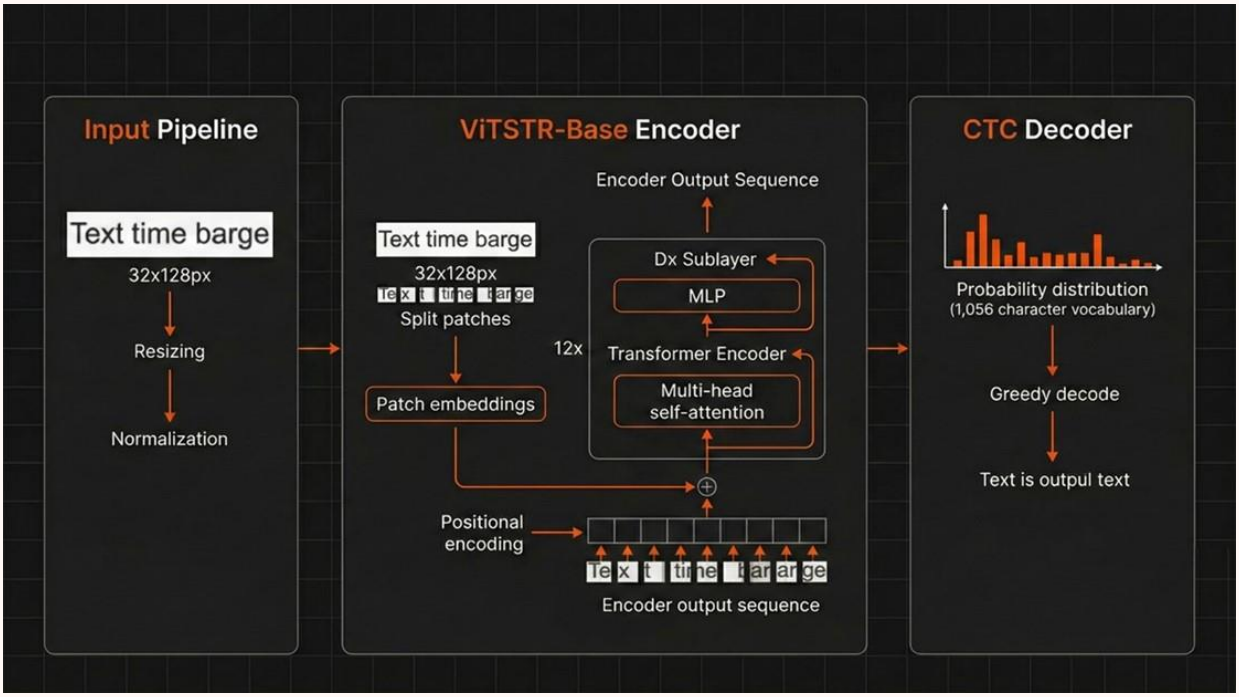

Figure 1: NE-OCR model architecture. A Vision Transformer encoder processes image patches and outputs character sequences via CTC decoding.

4.3. Architecture Configuration

Configuration	Value
Base Architecture	DocTR ViTSTR-Base
Parameters	86 million
Input Resolution	32 × 128 (H × W), RGB
Input Preprocessing	Resize → normalize to [0,1] → CHW tensor
Vocabulary Size	1,056 characters (index 0 = CTC blank)
Scripts Covered	4 (Latin, Devanagari, Bengali, Meitei Mayek)
Optimizer	AdamW
Learning Rate	1×10^{-4}
LR Schedule	CosineAnnealingLR
Batch Size	64
Gradient Clipping	1.0
Training Epochs	3
Loss Function	CTC Loss
Training Hardware	NVIDIA A40 48GB

Configuration	Value
Checkpoint Selection	Best validation Character Accuracy

Table 2 - NE-OCR architecture and training configuration.

5. Training Data

5.1. Data Construction Pipeline

Training data for NE-OCR was constructed from a combination of real and synthetically rendered text-image pairs, using language-specific configurations for each script. For languages where curated text collections were available, document and book data was incorporated directly. For languages with limited digitized resources, text samples were rendered into images using the Text Recognition Data Generator (TRDG) framework with Unicode-compatible fonts appropriate for each script.

Text samples span varying complexity levels - individual words, multi-word phrases, and short sentences. For synthetically rendered samples, visual diversity was introduced by randomizing font size, font weight, background style, text color, and character spacing. Images were further augmented with blur, noise, brightness variation, mild rotation, and compression artifacts to improve robustness across document types.

5.2. Data Sources

Training data was assembled from multiple sources depending on language and script. For Assamese, Hindi, and Meitei Bengali, data was drawn from the Mozhi Indic OCR dataset. For several Latin-script Northeast languages including Garo, Khasi, and Kokborok, curated book and document collections were used alongside rendered text. English training data was sampled from the MJSynth dataset (Jaderberg et al., 2014) to improve Latin character generalization. A supplementary Northeast place-name collection was included to improve recognition of proper nouns specific to the region.

Approximately 38% of the training data (514k samples) consists of real printed and scanned material from the Mozhi Indic OCR dataset (Assamese, Hindi, Meitei Bengali), English (MJSynth), and curated book/document scans.

Total training data across all language-script pairs is approximately 1.34 million images.

6. Experiments

6.1. Baselines

We compare NE-OCR against four baseline systems:

- **EasyOCR** - A widely deployed open-source OCR library supporting 80+ languages, using a CRNN-based recognition model (CNN + BiLSTM + CTC).
- **Tesseract 5** - The industry-standard open-source OCR engine, using LSTM-based recognition with language-specific trained data files.

- **TrOCR-large-printed** (microsoft/trocr-large-printed) - A transformer-based OCR model using a ViT encoder and language model decoder, achieving state-of-the-art results on English printed text benchmarks.
- **Chandra** - A vision-language model evaluated as a representative of the VLM paradigm for OCR on NE languages.

6.2. Main Results - Character Accuracy

Table 3 presents Character Accuracy (ChA%) for all five models across 12 language-script pairs. NE-OCR achieves the highest ChA on 9 pairs. Bold values indicate the best result per language-script pair.

Language	Script	NE-OCR (Ours)	EasyOCR	Tesseract 5	TrOCR (large-printed)	Chandra
		ChA% ↑	ChA% ↑	ChA% ↑	ChA% ↑	ChA% ↑
Assamese	*Bengali*	**97.46%**	32.25%	8.79%	0.80%	57.83%
Bodo	*Devanagari*	**83.38%**	82.65%	64.85%	1.85%	74.76%
English (anchor)	*Latin*	90.35%	68.91%	50.77%	88.87%	**91.30%**
Garo	*Latin*	93.52%	69.43%	69.90%	87.83%	**94.15%**
Hindi (anchor)	*Devanagari*	**97.69%**	49.54%	41.48%	1.27%	85.78%
Khasi	*Latin*	**98.85%**	77.78%	80.72%	93.22%	94.15%
Kokborok	*Latin*	**97.59%**	83.00%	78.76%	94.58%	96.19%
Meitei (Bengali script)	*Bengali*	**97.09%**	33.64%	7.30%	0.55%	48.34%
Meitei (Meitei Mayek)	*Meitei Mayek*	**95.56%**	2.50%	2.24%	2.45%	2.57%
Mizo	*Latin*	**95.96%**	67.62%	68.44%	84.58%	92.96%
Nagamese	*Latin*	**97.91%**	81.60%	78.05%	93.46%	97.60%
Nyishi	*Latin*	**94.50%**	69.56%	69.92%	87.23%	91.85%
Average	-	**94.99%**	**59.87%**	**51.77%**	**53.06%**	**77.29%**

Table 3: Character Accuracy (ChA%, higher is better) across 5 models and 12 language-script pairs. Bold = best per row. Avg row colored by model.

6.3. Character Error Rate

Language	Script	NE-OCR (Ours)	EasyOCR	Tesseract 5	TrOCR	Chandra
		CER% ↓	CER% ↓	CER% ↓	CER% ↓	CER% ↓

Language	Script	NE-OCR (Ours)	EasyOCR	Tesseract 5	TrOCR	Chandra
Assamese	*Bengali*	**2.15%**	63.27%	90.38%	105.29%	41.65%
Bodo	*Devanagari*	**10.93%**	12.04%	19.58%	97.84%	13.57%
English (anchor)	*Latin*	**8.60%**	25.15%	44.38%	66.21%	13.08%
Garo	*Latin*	4.13%	23.64%	22.63%	90.20%	**4.04%**
Hindi (anchor)	*Devanagari*	**1.86%**	47.90%	52.00%	104.04%	12.84%
Khasi	*Latin*	**0.84%**	17.25%	14.01%	85.85%	3.06%
Kokborok	*Latin*	**1.59%**	12.78%	16.20%	85.14%	2.05%
Meitei (Bengali script)	*Bengali*	**2.42%**	58.07%	90.08%	106.65%	61.12%
Meitei (Meitei Mayek)	*Meitei Mayek*	**2.71%**	95.64%	91.97%	98.24%	184.04%
Mizo	*Latin*	**2.54%**	24.77%	22.22%	84.12%	3.91%
Nagamese	*Latin*	**1.42%**	13.80%	15.59%	82.08%	2.02%
Nyishi	*Latin*	**3.88%**	23.77%	21.12%	85.47%	4.70%
Average	-	**3.59%**	**34.84%**	**41.68%**	**90.93%**	**28.84%**

Table 4: Character Error Rate (CER%, lower is better). Values above 100% indicate the model produces more characters than the ground truth.

6.4. Word Error Rate

Language	Script	NE-OCR (Ours)	EasyOCR	Tesseract 5	TrOCR	Chandra
		WER% ↓	**WER% ↓**	**WER% ↓**	**WER% ↓**	**WER% ↓**
Assamese	*Bengali*	**8.10%**	85.50%	94.75%	104.85%	82.40%
Bodo	*Devanagari*	34.68%	**30.90%**	41.52%	103.14%	45.46%
English (anchor)	*Latin*	**20.15%**	49.70%	74.55%	81.65%	28.70%
Garo	*Latin*	17.50%	51.26%	47.13%	103.68%	**17.18%**
Hindi (anchor)	*Devanagari*	**6.50%**	69.35%	73.71%	102.20%	26.97%
Khasi	*Latin*	**3.23%**	32.49%	24.89%	99.91%	10.40%
Kokborok	*Latin*	**6.86%**	32.61%	25.96%	99.89%	10.95%

Language	Script	NE-OCR (Ours)	EasyOCR	Tesseract 5	TrOCR	Chandra
Meitei (Bengali script)	*Bengali*	**10.00%**	88.35%	98.16%	109.60%	118.20%
Meitei (Meitei Mayek)	*Meitei Mayek*	**11.43%**	99.62%	106.46%	106.48%	184.48%
Mizo	*Latin*	**8.42%**	45.78%	44.06%	97.86%	12.65%
Nagamese	*Latin*	**5.41%**	30.54%	24.24%	99.16%	6.57%
Nyishi	*Latin*	**12.02%**	35.82%	36.31%	100.22%	16.11%
Average	-	**12.03%**	**54.33%**	**57.65%**	**100.72%**	**46.67%**

Table 5: Word Error Rate (WER%, lower is better). Values above 100% indicate more words generated than in the reference.

6.5. Inference Latency

Table 6 compare inference latency per image measured on an NVIDIA A40 48GB GPU. NE-OCR is the fastest system at 17.2ms per image, 9.7x faster than Tesseract 5 and 18.2x faster than Chandra.

Model	Latency (ms/image)	Relative to NE-OCR	Hardware
NE-OCR (Ours)	**17.2ms**	baseline	NVIDIA A40 48GB
EasyOCR	37.2ms	2.2× slower	NVIDIA A40 48GB
TrOCR-large-printed	92.1ms	5.4× slower	NVIDIA A40 48GB
Tesseract 5	166.1ms	9.7× slower	NVIDIA A40 48GB
Chandra	313ms	18.2× slower	NVIDIA A40 48GB
Latency measured on single-image inference, batch size=1, averaged over 10 runs on NVIDIA A40 48GB			

Table 6: Inference latency comparison. All measurements on NVIDIA A40 48GB GPU.

7. Analysis

7.1. Performance by Script

NE-OCR achieves its strongest results on Latin-script languages, with ChA values ranging from 90.35% (English anchor) to 98.85% (Khasi). The relatively lower English accuracy compared to dedicated English OCR systems reflects the absence of an English-specific fine-tuning stage; English is included primarily to improve Latin character generalization for Northeast languages. Among Bengali-script languages, Assamese achieves 97.46% and Meitei Bengali 97.09% ChA. On Devanagari, Hindi reaches 97.69% while Bodo achieves 83.38% - the lowest result across all language-script pairs.

Bodo presents the lowest Character Accuracy across all language-script pairs at 83.38%. Further investigation into Bodo-specific typographic characteristics, corpus quality, and character distribution is planned for NE-OCR v2.

7.2. Baseline Failure Modes

EasyOCR, Tesseract 5, and TrOCR-large-printed exhibit distinct failure modes on NE languages:

- **EasyOCR** achieves near-zero performance on Meitei Mayek (2.50% ChA), confirming that its training corpus contains negligible Meitei Mayek data. Bengali-script performance is also severely degraded for Assamese (32.25%) and Meitei Bengali (33.64%).
- **Tesseract 5** produces a WER of 106.46% on Meitei Mayek - inserting more words than are present in the ground truth - indicating that its internal language model is generating spurious insertions when encountering unseen character sequences.
- **TrOCR-large-printed** achieves high ChA on Latin-script languages (87–94%) but collapses on non-Latin scripts (<2% ChA on Assamese, Hindi, Bodo, Meitei Bengali, Meitei Mayek). Its CER and WER on these languages exceed 100%, reflecting the language model decoder generating English-like sequences regardless of visual input.

7.3. VLMs on NE OCR - A Preliminary Investigation

We evaluated two vision-language model systems as preliminary case studies: DeepSeek-OCR 2 (partial evaluation, 7 of 12 language-script pairs) and Chandra (full 12 pairs). Both systems demonstrate a characteristic failure mode distinct from traditional OCR errors: rather than producing incorrect character sequences, they hallucinate document structure.

Language	Ground Truth	DeepSeek-OCR 2 Output
Assamese	।	*<table><tr><td>Feature</td><td>Description</td></tr><tr><td>Object</td><td>Arrow</td></tr>...</table>*
Hindi	फल	*\[\pi \]*
English	Lube	*l1b*
Khasi	kiba thok.	*kîba thok.*

Table 7: Qualitative failure examples from DeepSeek-OCR 2 on NE language-script pairs.

The DeepSeek-OCR 2 failures shown in Table 7 illustrate the core problem: when presented with a single Assamese danda character (।), the model returns a multi-row HTML table describing image properties. When presented with the Hindi word for 'fruit' (फल), it outputs a LaTeX mathematical expression. These are not OCR errors - they are image understanding errors, where the model interprets the text image as a document or diagram rather than as a text recognition target.

Note on Chandra - Meitei Mayek Failure: Chandra achieves competitive results on Latin-script languages (Nagamese 97.60%, Kokborok 96.19%, Khasi 94.15%) but fails catastrophically on Meitei Mayek, producing a CER of 184.04% and WER of 184.48% - values exceeding 100% because the model generates substantially more characters and words than are present in the ground truth. This is

analogous to the hallucination pattern observed in DeepSeek-OCR 2 and reinforces the finding that general vision-language models are unable to handle unseen scripts without script-specific training data.

These findings suggest that VLMs, despite their impressive general capabilities, require script-specific training data or fine-tuning to perform reliably on unseen scripts. The NE-OCR approach - domain-specific training on native language corpora - remains the most effective strategy for the languages of Northeast India at this stage.

8. Model Access and Quick Start

NE-OCR model weights are publicly available on HuggingFace at huggingface.co/MWirelabs/ne-ocr under the CC-BY-4.0 license. The model can be loaded and used for inference using the DocTR library as follows:

```
# Install DocTR
pip install python-doctr

from doctr.io import DocumentFile
from doctr.models import ocr_predictor

# Load NE-OCR
model = ocr_predictor(
    reco_arch='MWirelabs/ne-ocr',
    pretrained=True
)

# Run on an image
doc = DocumentFile.from_images('your_image.jpg')
result = model(doc)
result.show()
```

Code snippet: Loading and running NE-OCR using the DocTR library.

9. Discussion

9.1. Unified vs. Per-Language Models

Training a single model across 4 scripts and 12 language-script pairs introduces vocabulary and script diversity that per-language models avoid. In practice, the shared ViT encoder benefits from cross-script training: Latin-script languages benefit from shared visual features across the 8 Latin-script pairs, and the transformer's attention mechanism learns script-level representations that generalize within each script family. The unified architecture is also operationally simpler - a single model checkpoint is deployable for all languages without language identification or routing logic.

9.2. Limitations

Several limitations should be noted. First, the benchmark test set is derived from the same data pipeline as training. Evaluation on independently sourced documents - scanned, photographed, and handwritten materials - remains future work. Second, Bodo presents a persistent accuracy gap (83.38% ChA) attributable to corpus size and font diversity constraints. Third, the model does not currently support text detection or layout analysis; it operates on pre-segmented text line crops. Fourth, the VLM analysis (Section 7.3) is preliminary and does not constitute a systematic benchmark.

9.3. Future Work

NE-OCR v2 is in preparation with expanded training data, improved augmentation pipelines, and document curation. Integration with text detection will provide an end-to-end document OCR pipeline for Northeast Indian languages. NE-OCR is one component of MWire Labs' broader language technology stack for the region.

10. Conclusion

We have presented NE-OCR, a unified OCR model for 10 Northeast Indian languages across 12 language-script pairs and 4 scripts, trained on approximately 1.34 million images. The model achieves a mean Character Accuracy of 94.99% across all language-script pairs - with a peak of 98.85% on Khasi - and outperforms EasyOCR, Tesseract 5, TrOCR-large-printed, and Chandra on 9 pairs. At 17.2ms per image on an A40 GPU, NE-OCR is the fastest system evaluated, 9.7 times faster than Tesseract 5.

Our analysis of VLM-based systems demonstrates that DeepSeek-OCR 2 and Chandra fail on unseen scripts by hallucinating document structure rather than producing recognition errors - a qualitatively distinct failure mode that highlights the continued importance of script-specific trained models for the languages of Northeast India.

Model weights and benchmark data are publicly available under the CC-BY-4.0 license. We expect this work to serve as a foundational resource for document digitization, accessibility technology, and language preservation efforts across Northeast India.

Acknowledgements

NE-OCR was developed by the research team at MWire Labs, Shillong, Meghalaya. We thank the open-source community for maintaining the language corpora, font collections, and OCR frameworks used in this work.

References

Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. ICDAR 2021, Springer, pp. 319–334.

Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. arXiv:1406.2227.

JaidedAI. (2020). EasyOCR: Ready-to-use OCR with 80+ languages supported. github.com/JaidedAI/EasyOCR.

Li, M., et al. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv:2109.10282.

Shi, B., Bai, X., & Yao, C. (2016). An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. IEEE TPAMI.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. ICDAR 2007.

Wei, H., Sun, Y., & Li, Y. (2025). DeepSeek-OCR 2: Visual Causal Flow. arXiv:2601.20552.

Mathew, M., Mondal, A., & Jawahar, C. V. (2025). Towards Deployable OCR Models for Indic Languages. In International Conference on Pattern Recognition (ICPR) (pp. 167–182). Springer.

Appendix A: Full Per-Language Results with Δ vs. Best Baseline

This appendix shows NE-OCR ChA versus the best-performing baseline for each language-script pair, along with the absolute improvement.

Language	Script	NE-OCR	EasyOCR	Tesseract 5	TrOCR	Chandra	Δ Best
Assamese	*Bengali*	**97.46%**	32.25%	8.79%	0.80%	57.83%	**+39.63%**
Bodo	*Devanagari*	**83.38%**	82.65%	64.85%	1.85%	74.76%	**+0.73%**
English (anchor)	*Latin*	**90.35%**	68.91%	50.77%	88.87%	91.30%	**-0.95%**
Garo	*Latin*	**93.52%**	69.43%	69.90%	87.83%	94.15%	**-0.63%**
Hindi (anchor)	*Devanagari*	**97.69%**	49.54%	41.48%	1.27%	85.78%	**+11.91%**
Khasi	*Latin*	**98.85%**	77.78%	80.72%	93.22%	94.15%	**+4.70%**
Kokborok	*Latin*	**97.59%**	83.00%	78.76%	94.58%	96.19%	**+1.40%**
Meitei (Bengali script)	*Bengali*	**97.09%**	33.64%	7.30%	0.55%	48.34%	**+48.75%**
Meitei (Meitei Mayek)	*Meitei Mayek*	**95.56%**	2.50%	2.24%	2.45%	2.57%	**+92.99%**
Mizo	*Latin*	**95.96%**	67.62%	68.44%	84.58%	92.96%	**+3.00%**
Nagamese	*Latin*	**97.91%**	81.60%	78.05%	93.46%	97.60%	**+0.31%**
Nyishi	*Latin*	**94.50%**	69.56%	69.92%	87.23%	91.85%	**+2.65%**

Table A1: NE-OCR ChA vs. all baselines with improvement over best baseline (Δ Best).